Journal of Financial Stability 65 (2023) 101099

Contents lists available at ScienceDirect

Journal of Financial Stability

journal homepage: www.elsevier.com/locate/jfstabil

Forecasting Stock Market Crashes via Machine Learning☆

Hubert Dichtl, Wolfgang Drobetz, Tizian Otto *

Faculty of Business Administration, University of Hamburg, Moorweidenstrasse 18, 20148 Hamburg, Germany

ARTICLE INFO

JEL classification codes:
G11
G12
G14
G17

Keywords:
Extreme event prediction
Stock market crashes
Machine learning
Active trading strategy

ABSTRACT

This paper uses a comprehensive set of predictor variables from the five largest Eurozone countries to compare the performance of simple univariate and machine learning-based multivariate models in forecasting stock market crashes. In terms of statistical predictive performance, a support vector machine-based crash prediction model outperforms a random classifier and is superior to the average univariate benchmark as well as a multivariate logistic regression model. Incorporating nonlinear and interactive effects is both imperative and foundation for the outperformance of support vector machines. Their ability to forecast stock market crashes out-of-sample translates into substantial value-added to active investors. From a policy perspective, the use of machine learning-based crash prediction models can help activate macroprudential tools in time.

1. Introduction

Predicting extreme events out-of-sample early and accurately, i.e., predicting all financial crises which actually happened without crying wolf, is notoriously difficult. As Fouliard et al. (2021) note, the ability of existing early-warning models to predict turning points or nonlinear and interactive phenomena out-of-sample is still limited. Examples of extreme events that occur infrequently and irregularly are banking crises, sovereign debt crises, private debt crises, and currency crises (Kaminsky and Reinhart, 1999; Reinhart and Rogoff, 2011; Jordà et al., 2017; Baron et al., 2021). Such rare and far between crisis events, causing substantial economic, political, and social costs, are obviously a source of constant concern. For example, Laeven and Valencia (2020) document that the average cumulative aggregate output loss in a banking crisis (computed as the deviation of the actual gross domestic product from its trend) is around 20% over the length of the crisis, which is, on average, two years.

In our empirical analysis, we put the focus on forecasting stock market crashes. It is surprising that most of the macrofinance literature ignores such crises, despite stock markets being an important indicator of the expected economic development in the future and a means of wealth storage for both institutional and retail investors.[1] A potential reason could be that stock market crashes originate from other financial crises, such as banking crises, debt crises, or currency crises, and not vice versa. If a sudden and sharp drop in aggregate stock prices is merely considered a reflection of expected economic shrinkage or impediment caused by the underlying crises, stock market crashes are not identified as isolated financial crises that cause economic contractions or even recessions. Nevertheless, Barro and Ursúa (2017) document that stock market crashes are informative about the prospects for economic depressions. Crashes are more frequent than depressions, but the largest depressions are particularly likely to be accompanied by preceding crashes. They conclude that, in the absence of a crash, the occurrence of a depression is highly unlikely. Huang and Chang (2022) document that stock market crises have an isolated negative impact on economic growth, even after controlling for the consequences of other types of financial crises. They can hurt consumption and investment activities (e.g., through an increase in firms' equity cost of capital), resulting in a negative impact on economic growth. These findings delineate an isolated negative impact of stock market crises on the real economy. An accurate model for predicting stock market crashes out-of-sample can help activate macroprudential policy tools in time, aiming at increasing the resiliency of the financial system as a whole and mitigating the

☆ Acknowledgments: We thank two anonymous referees, Mikhail Chernov, Valentin Haddad, Barney Hartman-Glaser, Kay Giesecke, Lisa Goldberg, Iftekhar Hasan (the editor), Bernard Herskovic, Bryan Kelly, Martin Lettau, Lars A. Lochstoer, Tyler Muir, Andreas Neuhierl, Stavros Panageas, Markus Pelger, Terrance Odean, Michael Weber, Ivo Welch, and Dacheng Xiu for helpful comments.

* Corresponding author.
E-mail addresses: hubert.josef.dichtl@uni-hamburg.de (H. Dichtl), wolfgang.drobetz@uni-hamburg.de (W. Drobetz), tizian.otto@uni-hamburg.de (T. Otto).

[1] For example, in their surveys of financial crises, Claessens and Kose (2013), Reinhart and Reinhart (2015), and Gorton (2018) do not explicitly mention stock market crashes when defining and discussing different crisis events.

https://doi.org/10.1016/j.jfs.2022.101099
Received 30 July 2021; Received in revised form 6 December 2022; Accepted 7 December 2022
Available online 19 December 2022

imminent costs of financial crises.

Both the actual occurrence and the mere possibility of a stock market crash also loom large for equity investors. Goetzmann et al. (2017) document that the average individual investor, responding to a survey, estimates that a catastrophic stock market crash within the next six months has a 20% chance of happening, which is much higher than implied by the historical frequency of such events. Moreover, because crashes influence expectations about future returns, such fears can reinforce the cyclical nature of stock markets and impose a threat on financial and macroeconomic stability. For example, institutional investors allocate equity procyclically (Ang et al., 2014; Goyal et al., 2015; Jones, 2016), and subjective crash probabilities of individual investors are negatively associated with mutual fund flows (Goetzmann et al., 2017). Therefore, an accurate model for predicting stock market crashes out-of-sample enables active investors to implement tactical portfolio adjustments. Falsely classifying a crash month as a non-crash month is costlier in terms of investment performance than falsely classifying a non-crash month as a crash month due to the asymmetry in stock returns (Dichtl et al., 2016). As a result, the economic costs of forecast errors, most importantly, missing a crash and staying invested in the stock market, are immediately reflected in the performance of market timing and switching strategies.[2]

However, identifying predictors for extreme stock market events is exceptionally difficult. Because they occur seldom, they are likely to exhibit unknown and time-varying patterns of nonlinearity and interactions in the relationship between predictor variables and probabilities of occurrence (McPhillips et al., 2018). Furthermore, scarce occurrences result in highly imbalanced datasets that can hamper robust modeling. Both theoretical findings (Wolpert and Macready, 1997) and empirical evidence (Fernández-Delgado et al., 2014) suggest that machine learning algorithms are particularly well suited to overcome these methodological challenges. For example, recent studies use machine learning techniques for banking crisis prediction (Tanaka, Kinkyo, and Hamori, 2016; Alessi and Detken, 2018; Beutel et al., 2019; Bluwstein et al., 2020; Samitas et al., 2020; Fouliard, Howell, and Rey, 2021). While there is also growing research that uses machine learning-based methods in the empirical asset pricing literature (Freyberger, Neuhierl, and Weber, 2020; Gu et al., 2020; Drobetz and Otto, 2021), attempts to apply them for mere directional forecasts (rather than level forecasts) and the prediction of large and sudden stock market declines are still scant. To date, Chatzis et al.'s (2018) study is so far the only one that systematically addresses the problem of forecasting future stock market crashes via machine learning.

The theoretical asset pricing literature establishes a potential link between the predictability of stock market crashes and the concept of bubbles. Asset price bubbles are defined as the deviation of an asset's market price from its fundamental value because current owners believe they are able to resell the asset at an even higher price later (Brunnermeier, 2009). In his Nobel Lecture, Fama (2014) argues that for a bubble to exist on stock markets, irrationally strong price increases must imply a predictably strong price decline (indicative of a bubble's burst). Greenwood et al. (2019) show that, consistent with the weak-form market efficiency, longer price run-ups do not predict low subsequent returns. Nevertheless, they predict a heightened probability of a crash in the near future. Kaminsky and Reinhart (1999) conclude that what may (or may not) be an asset price bubble's burst is most likely to be observed ahead of multidimensional financial crisis events.[3] These results support univariate valuation measures to predict stock market crashes such as Campbell and Shiller's (1988a, 1988b) model based on the earnings-to-price ratio, the bond-stock earnings yield differential (BSEYD) model (Ziemba and Schwartz, 1991; Lleo and Ziemba, 2017), or the Fed model (Asness, 2003; Estrada, 2006; Maio, 2013).

In our empirical setting, we refer to the concept of semi-strong market efficiency and, extending Greenwood et al.'s (2021) study, examine whether any information in addition to the attributes of the price run-up helps predict stock market crashes. Following Chatzis et al. (2018), we define crash months as months in which the stock market return is below the 5^{th} percentile based on the historical distribution over a ten-year rolling window. Based on this definition, we provide a framework for a systematic and consistent comparison 1) across univariate crash prediction models and 2) between univariate and multivariate crash prediction models. We use a comprehensive set of price-based, fundamentals-based, sentiment-based, and macroeconomic predictor variables from the five largest Eurozone countries. As in Lleo and Ziemba (2017, 2019), we construct univariate models using percentile-based thresholds for each predictor in isolation. Multivariate approaches are machine learning-based and incorporate information from multiple different predictors simultaneously. We use logistic regressions and more sophisticated machine learning techniques represented by support vector machines (SVMs), which are a well-established and popular classification method (Vapnik, 1998). In a robustness test, we use other machine learning classifiers such as two tree-based models (random forests and gradient boosted regression trees) and neural networks.

Our results show that, in terms of statistical predictive performance, an SVM-based crash prediction model outperforms a random classifier and is superior to the average univariate benchmark as well as a multivariate logistic regression model. We demonstrate, based on several instructive examples in the run-up to stock market crashes, that incorporating nonlinear and interactive effects is both imperative and foundation for the outperformance of SVMs. Their ability to forecast stock market crashes out-of-sample translates into substantial value-added to active investors under realistic trading assumptions. In a larger sense, using machine learning techniques enables us to uncover robust statistical relationships in the underlying economic conditions that precede stock market crashes. For example, abnormally low stock market returns in the current month do not necessarily point towards an even stronger stock market correction in the next month, *unless* other economic indicators are simultaneously falling out of their normal ranges (e.g., the yield curve flattens or even reverses). Our results thus confirm Lopez de Prado's (2020) insight that machine learning should not prematurely be regarded as a black box.

The question *why* stock market crashes happen at all is clearly beyond the scope of our empirical analysis. We are unable to answer the more fundamental question whether the stock price dynamics we observe are attributable to changes in expected returns or discount rates driven by fundamental variables such as conditions affecting profits, risk, and risk premiums (Pastor and Veronesi, 2006, 2009), or are the result of bursting or deflating bubbles, which can either be of rational or behavioral origin (Brunnermeier, 2009; Brunnermeier and Oehmke, 2013; Scherbina and Schlusche, 2014). Nevertheless, our results have several important implications: First, they are valuable for long-term investors, whose risk-adjusted performance improves when they are able to forecast large stock market corrections. Second, stock markets are a leading indicator of the real economy (Fischer and Merton, 1984).

[2] Market timing refers to moving investment funds in or out of a particular financial market (or asset class), and market switching to moving them between different financial markets (or asset classes), based on predictive methods. If an investor can predict when financial markets will go up and down, market timing and switching trades can turn these price movememts into profits.

[3] Many studies document that rapid rises of aggregate stock prices are predictive of financial crises, in particular, when accompanied by high credit growth (Schularick and Taylor, 2012; Aliber and Kindleberger, 2015; Greenwood et al., 2021).

Predicting stock market crashes is distinct – albeit clearly not fully separable – from predicting other financial crises, and it can help forecast financial fragility and deteriorating macroeconomic outcomes. Through these channels, our results are related to financial and macroeconomic stability.

Our main objective is to examine whether machine learning-based approaches outperform simple univariate benchmarks as well as a multivariate logistic regression model in predicting future stock market crashes and, if yes, why. In our empirical analysis, we select a framework that resembles the realistic trading patterns of European institutional investors. They typically restructure their portfolios on a monthly basis and are often required to maintain investments in different Eurozone countries.

While Chatzis et al.'s (2018) study is clearly related to ours, we expand it in various dimensions: First, we do not only compare the predictive performance of crash prediction models from a statistical perspective, but also contrast their economic profitability through the lens of an active investor. To this end, we analyze the performance of market timing and switching strategies.[4] Second, rather than looking at the global stock market, we work with data from the five largest Eurozone countries, which allows predicting crashes at a more granular level and even assists active investors in tactically allocating their funds across countries over time. Third, we do not limit our crash prediction models to price-based market data. Against the backdrop of semi-strong market efficiency, we use a comprehensive set of price-based, fundamentals-based, sentiment-based, and macroeconomic predictor variables that have been shown to explain and forecast subsequent stock market crashes. We also incorporate predictors that have not yet been widely used in the literature, despite their likely predictive power, such as the metric from Chow et al. (1999) and Kritzman and Li (2010) that captures the financial turbulence on stock markets.

Fourth, extending Chatzis et al.'s (2018) study, we address the black box issue in predicting stock market crashes by investigating the characteristics and functioning scheme of machine learning techniques. Inspecting changes in the inherent model complexity over time, we observe strong time variation in the degree of model complexity and a co-movement with the current-month stock market variance, indicating that high-volatility periods are more difficult to predict. We also decompose predictions into the contributions of individual variables using relative variable importance metrics.[5] The most influential predictors are based on exchange rate trends, returns on stock, oil, and gold markets, and variables reflecting current stock market risk. In contrast, information from bond markets seem to be less relevant.

We show that there is no single variable or small subset of variables that always and reliably precedes stock market crashes with extreme values, suggesting substantial time variation in the predictive ability of any single variable. Because multivariate crash prediction models are capable of incorporating multiple predictor variables simultaneously, they should be advantageous over their univariate counterparts. Supporting this view, we show that only a few predictors in the univariate setting, e.g., valuation metrics such as the earnings-to-price ratio or the dividend-to-price ratio, deliver significant predictive performance over the full sample period. However, in contrast to our multivariate approaches, their predictive ability is often not evenly distributed over the sample period but strictly limited to shorter subperiods, e.g., the dotcom bubble.

Finally, given logistic regressions as our linear multivariate benchmark, we explore patterns of nonlinear and interactive effects in the relationship between predictor variables and the estimated crash probabilities, which are likely to be inherent to extreme events. Our results show that incorporating these effects is pivotal for the superior predictive performance of SVMs relative to their linear counterpart.

The remainder is organized as follows: Section 2 reviews the literature on crash prediction and machine learning. Section 3 describes our dataset. Section 4 analyzes the economic conditions preceding crashes. Section 5 introduces univariate and multivariate crash prediction models. Section 6 assesses the predictive performance of these models and analyzes model complexity and variable importance as well as nonlinear and interactive effects. Section 7 concludes.

2. Literature review

According to Chatzis et al. (2018), stock market crashes are defined as periods during which stock market returns are abnormally low relative to their historical distribution. In this light, the literature applies statistical models to predict crashes along two strands: univariate models and multivariate models.

Univariate crash prediction models are the simplest. As in Ziemba et al. (2017), we classify them into two subcategories based on methodology. The first subcategory refers to *stochastic models*, which are probabilistic representations of stock markets. These models are based on the idea that stock market states are characterized by price-based metrics such as return, volatility, and (auto-)correlation (Ang and Bekaert, 2004; Ang and Timmermann, 2012; Neely et al., 2014). Applying sophisticated algorithms, they aim to identify patterns in historical stock market prices such as regime shifts (Bulla et al., 2011; Nystrup et al., 2015) or change points (Shiryaev et al., 2014).

The second subcategory consists of models restricted to only one variable, which apply a heuristic that mimics our crash definition, meaning that abnormally high or low values of a predictor relative to its historical distribution generate crash signals. These models are distinguished by the predictor variable under investigation. However, they have in common that economic theory postulates some degree of predictive ability for the specific variable used.

Fundamentals-based models link stock market movements to changes in the firms' fundamentals or macroeconomic environment, and consider a wide range of possible valuation metrics. Examples include the earnings-to-price ratio (Campbell and Shiller, 1988a, 1988b) and the bond-stock earnings yield differential (BSEYD) metric (Ziemba and Schwartz, 1991; Lleo and Ziemba, 2017), which relates the yield on ten-year government bonds to the inverse of the price-to-earnings ratio.

Sentiment-based models assume that stock market movements are caused by changes in the overall economic sentiment (Shiller, 2003; Baker and Wurgler, 2006, 2007). For example, according to Billingsley and Chance (1988), Copeland and Copeland (1999), Whaley (2000), Bandopadhyaya and Jones (2008), and Goetzmann et al. (2017), put-call ratios or implied volatilities, among others, can be used as sentiment-based predictors.

Multivariate crash models are more complex. They are devised to 1) incorporate a large set of variables simultaneously and 2) consider both nonlinearity and interactions in the relationship between predictors and subsequent stock market downturns. Leung et al. (2000) conduct a comparative analysis of simple classification-based methods (including linear discriminant analysis and logistic regressions) to predict the direction of stock market movements and provide evidence for substantial predictive ability. Chatzis et al.'s (2018) study is so far the only one that

[4] We choose a monthly setting and re-estimate our machine learning-based models each month, rather than using daily data and fitting machine learning classifiers based on only a single sample split. Therefore, our approach allows to better account for the time variation in the predictive ability of single variables. This is important in light of the Lucas (1976) critique, stressing that the structure of econometric models changes across different economic environments and crisis times (see Section 4.2).

[5] We examine the importance of each predictor in a given crash prediction model over the full sample period, but we also scrutinize the time variation in each predictor variable's importance. This approach reveals whether 1) some predictors are uninformative during the entire sample period, 2) they lead to a permanent deterioration in a forecast's signal-to-noise ratio, and 3) they should be removed from the set of baseline variables.

systematically addresses the problem of forecasting future stock market crashes via machine learning.[6] They also find significant predictive power of multivariate crash prediction models and conclude that machine learning techniques (including SVMs, tree-based models, and neural networks) outperform simpler approaches. However, both studies only focus on price-based data from stock, bond, or currency markets.

Several studies that aim to forecast the direction of future stock market movements incorporate additional, non-price-based information to increase the predictive performance. For example, Choudhry and Garg (2008) apply a hybrid machine learning system. While also using several variables based on technical analysis, they propose that including fundamentals- and sentiment-based variables improves classification accuracy. Both Lee et al. (2019) and Ren et al. (2019) use various machine learning techniques to highlight the incremental predictive power of sentiment-based and financial network indicators. Earlier studies (Huang et al., 2005; Shao and Lunetta, 2012) provide evidence that SVMs outperform other classification methods (including random walk models, quadratic discriminant analysis, tree-based models, and neural networks), particularly when working with small sample sizes. Therefore, the literature often only focuses on how different versions of SVMs predict the direction of future stock market movements.[7]

3. Data

Our data come from Refinitiv and comprise market and fundamental data from stock, bond, commodity, and forex markets as well as sentiment and macroeconomic indicators. They are collected on a monthly basis and, if currency-related, denominated in Euro. Our sample includes the five largest Eurozone countries by gross domestic product as of December 2019, i.e., Germany, France, Italy, Spain, and the Netherlands. However, we omit country observations with missing values on stock market returns or at least one of the twenty-eight predictor variables used in the empirical analysis. This shrinks our sample period to January 1990–December 2020. To calculate excess returns, we use the three-month FIBOR or EURIBOR rate, whichever is available, scaled to the one-month horizon, as the risk-free rate.

We follow Welch and Goyal (2008), Lleo and Ziemba (2017), Chatzis et al. (2018), Neely et al. (2014), and Bluwstein et al. (2020), among others, and construct a comprehensive set of variables that have been shown to predict future stock market returns. In line with Chatzis et al. (2018), who differentiate between local and global financial crises, we refer to stock market crashes that sporadically occur in a single country as idiosyncratic (local) and those that simultaneously occur in multiple/all countries as systematic (global). We assume that local crashes are driven mainly by changes in country-specific economic conditions, while global crashes originate from supranational forces. As shown in Table 1, each country's dataset thus consists of two categories of predictors: *supranational variables*, e.g., based on gold, oil, or foreign exchange markets, and *country-specific variables*, e.g., based on domestic stock and bond markets.

We further include variables that have not yet been widely used in the literature to predict stock market crashes, although they contain information that is likely to add incremental predictive performance. Ferrer et al. (2016) propose that the consumer confidence index (*cci*) serves as an indicator of a country's overall economic sentiment, and possesses predictive power for future stock market meltdowns. Furthermore, in order to be meaningful, crash predictions must include information on whether the current stock market return and risk metrics are abnormal in comparison with their own past. To this end, we adapt the distance ratio of Avramov et al. (2020), which is defined as the ratio between short- and long-run moving averages of stock prices and has been shown to possess predictive power for the cross section of stock-level expected returns. Adjusted for our empirical setting, we apply this ratio to both stock market returns and variances (*mrat* and *svrat*, respectively).

Finally, we include three variables that explicitly capture the systematic component of stock market crash risk. First, we use the German composite indicator of systematic stress (*ciss*) provided by the European Central Bank (Holló et al., 2012). High values indicate high systematic crash risk. Second, we incorporate the financial turbulence metric (*ft*) from Chow et al. (1999) and Kritzman and Li (2010), which identifies conditions under which stock markets behave atypically given their historical patterns. Uncharacteristic behavior (such as extreme stock market returns, decoupling of correlated stock markets, or convergence of uncorrelated stock markets) is quantified based on the Mahalanobis (1936) distance measure. Low values reflect a low similarity with historical patterns, indicating high systematic crash risk. Third, we include the absorption ratio (*ar*) introduced by Kritzman et al. (2011), which applies a principal component analysis to the historical stock market returns of the five sample countries. It is defined as the fraction of the total variance absorbed by the first two eigenvectors. High values display high stock market fragility, corresponding to high systematic crash risk.

Table 1 provides the definitions and time series means for the twenty-eight predictors that serve as the starting point for our empirical analysis. One important caveat is that many of the predictors are constructed similarly or incorporate similar information, which leads to relatively high correlations. However, according to Lewellen (2015), any resulting multicollinearity is not a major concern because we are mostly interested in the overall predictive power of machine learning-based crash prediction models, rather than the marginal effects of each single predictor.

4. Characteristics of stock market crashes

Our main objective is assessing the ability to forecast future stock market crashes. Any crash prediction model, whether univariate or multivariate, consists of two components: 1) a binary *crash indicator* CI_{t+1}, which equals 1 when a substantial stock market downturn occurs during month $t+1$, and 0 otherwise, and 2) a binary *crash signal* $CS_{t+1|t}$, which equals 1 if the model, incorporating all information available at the end of month t, expects a stock market crash to occur during month $t+1$, and 0 otherwise.

4.1. Stock market crashes over time

We follow Chatzis et al. (2018) in defining stock market crashes.[8] We derive the binary *crash indicator* CI_{t+1}, separately for each country, by

[6] Ohana et al. (2021) also use machine learning techniques to predict stock market crashes. However, they do not offer a systematic framework to assess their predictive ability because the authors focus on only one specific stock market crash (the S&P 500 meltdown in March 2020).

[7] Yu et al. (2005) and Yu et al. (2009) implement an evolving least squares SVM with a Genetic Algorithm (GA) for both feature selection and parameter tuning. Similarly, Choudhry and Garg (2008) and Khatibi et al. (2011) hybridize SVM and GA, while Ni et al. (2011) add a fractal dimension-based feature selection.

[8] Similar to other extreme events, stock market crashes lack uniform definition. For example, Goetzmann et al. (2017) define crashes as substantial one-day drops in a stock market, while Goetzmann and Kim (2018) define them as events for which the stock market declines by more than 50% over one year. In both examples, the crash indication depends strongly on the return threshold, which is commonly chosen by discretion and kept constant over time. During high- or low-volatility periods, however, it may be more appropriate to consider higher or lower thresholds, respectively. Chatzis et al. (2018) use a percentile-based approach of crash indication, which reduces the need for discretionary selection and allows for time variation in the threshold. In particular, they link stock market crashes to extreme negative return events (based on the historical return distribution) that were likely caused by substantial changes in the underlying economic conditions.

Table 1
Descriptive statistics.

#	Predictor	Definition	Mean				
Supranational predictors			Overall				
1	ar	Absorption ratio	0.93				
2	ciss	German composite indicator of systematic stress	0.15				
3	dfy	Default spread [%]	0.08				
4	ft	Financial turbulence metric	5.46				
5	ret_exr	One-month percentage change in U.S. dollar-to-Euro exchange rate [%]	0.11				
6	ret_exr_ann	One-year percentage change in U.S. dollar-to-Euro exchange rate [%]	0.80				
7	ret_gold	One-month gold market excess return [%]	0.70				
8	ret_gold_ann	One-year gold market excess return [%]	8.67				
9	ret_oil	One-month oil market excess return [%]	0.74				
10	ret_oil_ann	One-year oil market excess return [%]	8.08				
11	tbl	Three-month FIBOR or EURIBOR rate, scaled to the one-month horizon [%]	0.13				
12	tds	TED spread (between three-month LIBOR rate denoted in U.S. dollar and three-month U.S. T-bill rate) [%]	0.03				
13	tms	Term spread (between yield on ten-year German government bonds and three-month FIBOR or EURIBOR rate, scaled to the one-month horizon) [%]	0.07				
Country-specific predictors			BD	FR	IT	ES	NL
14	bm	Book-to-market ratio	0.62	0.59	0.82	0.60	0.63
15	bseyd	Bond-stock earnings yield differential [%]	-4.47	-3.60	-2.77	-3.34	-3.86
16	cci	Consumer confidence index [factor 1e4]	-0.16	-1.50	-0.91	-2.72	-1.20
17	dp	Dividend-to-price ratio	0.41	0.34	0.30	0.31	0.37
18	ep	Earnings-to-price ratio	0.07	0.06	0.07	0.07	0.07
19	ir	One-month percentage change in yield on ten-year government bonds [%]	0.20	0.23	0.31	0.29	0.22
20	ir_ann	One-year percentage change in yield on ten-year government bonds [%]	2.49	2.82	3.77	3.54	2.69
21	mrat	Distance ratio (between one-month and one-year stock market return)	1.02	1.03	1.01	1.02	1.02
22	ntis	Net equity expansion [%]	0.75	1.65	1.46	2.79	-0.47
23	ret_stock	One-month stock market excess return	0.39	0.43	0.24	0.34	0.36
24	ret_stock_ann	One-year stock market excess return	5.22	6.30	2.94	4.22	5.05
25	svar	One-month stock market variance (mean of squared daily stock market returns) [factor 1e4]	1.52	1.65	1.84	1.72	1.56
26	svar_ann	One-year stock market variance (mean of squared daily stock market returns) [factor 1e4]	1.49	1.61	1.80	1.68	1.54
27	svrat	Distance ratio (between one-month and one-year stock market variance)	1.04	1.06	1.06	1.05	1.06
28	to	Turnover [€ bill.]	7.94	92.41	61.15	42.36	44.35

This table gives the definitions and time series means for the twenty-eight supranational and country-specific predictors used in the empirical analysis. The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, BD, France, FR, Italy, IT, Spain, ES, and the Netherlands, NL) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

comparing the stock market return $ret_{stock,t+1}$ during month $t+1$ with a historical return-based threshold $HRT_{stock,t+1|t}$ for month $t+1$, calculated at the end of each month t. This threshold is defined as the 5^{th} percentile based on the historical stock market return distribution over a ten-year rolling window.[9] Accordingly, a crash during month $t+1$ is indicated if the respective stock market return falls below the calculated threshold:

$$CI_{t+1} = \begin{cases} 1 \text{ if } ret_{stock,t+1} - HRT_{stock,t+1|t} < 0 \\ 0 \text{ otherwise} \end{cases}. \quad (1)$$

Panel A of Fig. 1 illustrates our procedure for Germany. The red line represents the historical return-based threshold. Non-crash months (stock market returns above the threshold) are marked with a black unfilled circle, while crash months (stock market returns below the threshold) are marked with a red unfilled circle. Panel B of Fig. 1 visualizes the aggregate number of crash occurrences across the five sample countries over time. It reveals three periods during which crashes occured simultaneously in nearly all countries (grey-shaded areas) and periods with only a sporadic crash in single countries. This phenomenon points towards systematic and idiosyncratic components of stock market crash risk, which is in line with Chatzis et al.'s (2018) notion of global and local stock market crashes.[10]

[9] The identified patterns are similar for alternative definitions of stock market crashes, e.g., using the 10^{th} percentile based on the historical stock market return distribution over a ten-year rolling window as the threshold.

[10] Because crash prediction models are required to identify both systematic and idiosyncratic types, we incorporate supranational and country-specific variables and do not distinguish between these two types in our empirical analysis.

4.2. Economic conditions preceding stock market crashes

Anecdotal evidence links the three periods to economically distinct crash episodes, namely the *dotcom bubble*, the *Global Financial Crisis*, and the *Covid-19 pandemic*. We now examine the commonalities and differences among the economic conditions that preceded these crash episodes. For the sake of brevity, we report the descriptive statistics illustratively for Germany and only a single crash month during each of the three periods (September 2002, January 2008, and March 2020).[11] As specific numbers (in levels) are less informative than whether these numbers are large or small relative to their historical distributions, we apply a detrending and scaling procedure.

In a first step, for each predictor, we compute the detrended values as the residuals of an ordinary least squares regression of the raw values on a time dummy. We use a five-year rolling window and all information up to the respective point in time. In a second step, following Kelly et al. (2019) and Freyberger et al. (2020), we map the detrended values into the $(-1,+1)$ interval. Positive numbers close to $+1$ indicate abnormally high values, while negative numbers close to -1 indicate abnormally low values. Hypothetically, there could be two different findings: First, prior to each crash, a single variable or small subset of variables always has abnormal numbers, which emphasizes stable predictive power. Second, the variables with abnormal figures differ across crashes, indicating time variation in their predictive ability.

[11] For the sake of brevity, we omit the six one-year predictors, because the findings are similar to their one-month equivalents, and thus only show the results for the remaining twenty-two variables to assess the crash characteristics.

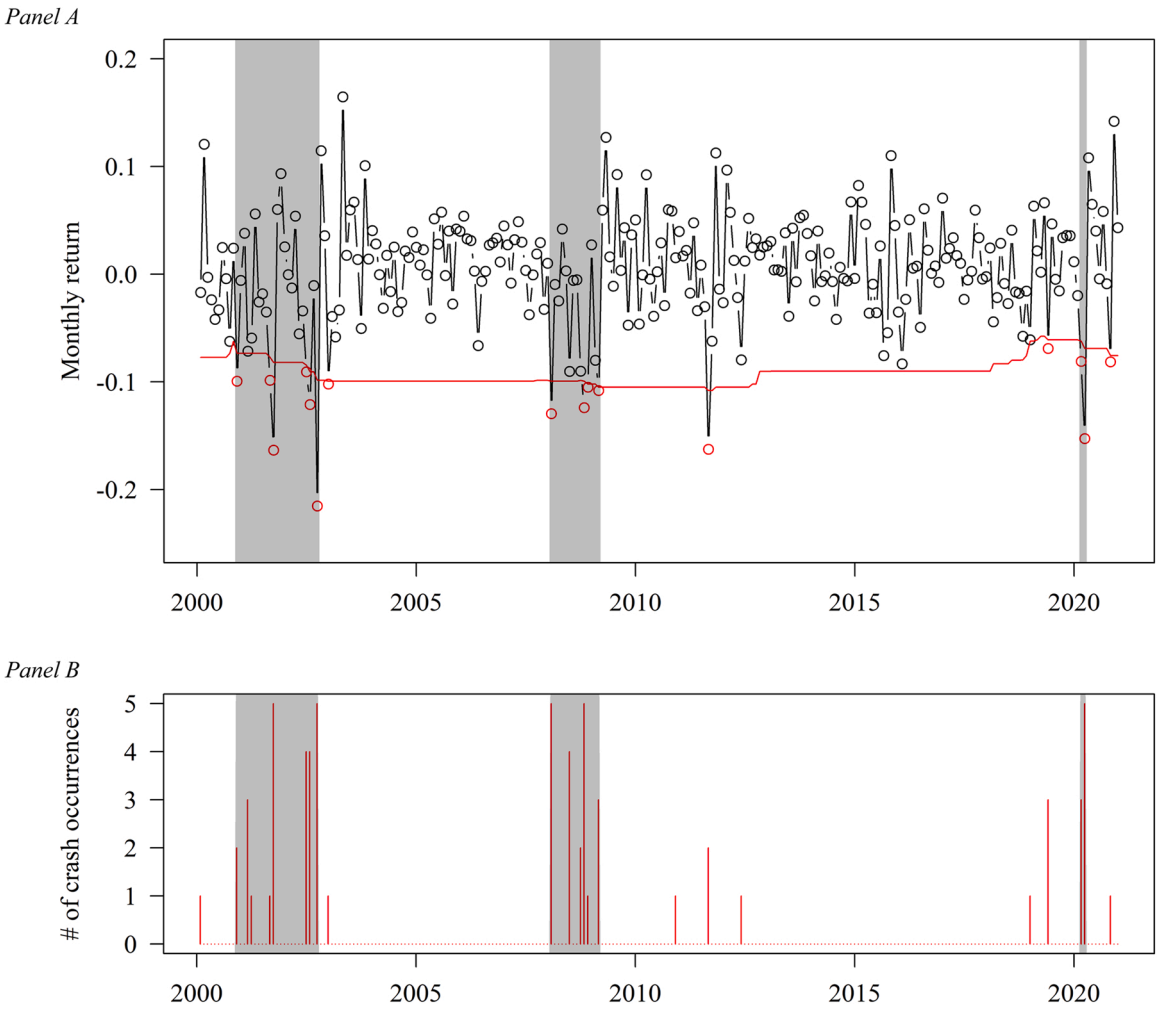

Fig. 1. Stock market crashes over time. This figure presents the indicated stock market crashes during the January 2000–December 2020 out-of-sample period. Panel A visualizes the procedure to obtain the binary *crash indicator* CI_{t+1}, which equals 1 when a substantial stock market downturn occurs during month t + 1, and 0 otherwise, illustratively for Germany. The red line represents the historical return-based threshold. *Non-crash months* (stock market returns above the threshold) are marked with a black unfilled circle, while *crash months* (stock market returns below the threshold) are marked with a red unfilled circle. Panel B adds the aggregate number of crash occurrences across the five sample countries over time. The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

Fig. 2 presents the respective Kiviat charts illustratively for Germany and the crashes in September 2002, January 2008, and March 2020. Abnormally high values are located close to the border, whereas abnormally low values close to the center. The crashes have in common that the polygons are close to the border or center for many predictors, suggesting that the economic conditions preceding these stock market crashes were extreme. However, the Kiviat charts differ substantially across crashes, showing that there was no single variable or small subset of variables that always and reliably preceded these stock market crashes with extreme values. We conclude that multivariate crash prediction models, which incorporate a comprehensive set of variables simultaneously, are potentially advantageous over approaches that consider only a single predictor or small set of predictors.[12]

These observations are related to the Lucas (1976) critique. Economic relationships change in response to changes in the expectations of market participants and their endogenously determined decisions. McLean and Pontiff (2016) document that investors learn about mispricing from academic publications. Many of the stock market predictors uncovered in the finance research lost their predictive power soon after they were published in academic journals. There may also be an inherent self-defeating property in crash prediction: Variables that performed well in historical samples may lose their predictive ability for future crashes. To mitigate this problem, we move away from univariate models, where the predictive ability of the sole predictor may vanish over time, and incorporate a comprehensive set of predictor variables simultaneously. More importantly, we recursively refit the machine learning models each month. This time-consuming calibration approach allows regular updates of the historical crash patterns and reduces the necessary stability in the economic relationships to only a short time period, namely one month. It further takes into account any changes in the relative importance of each predictor variable and, at any specific point in time, puts the focus on those predictors that are most informative.

[12] The results for the remaining crashes and sample countries (unreported) are similar and underline that crash characteristics vary substantially in both time series and cross section.

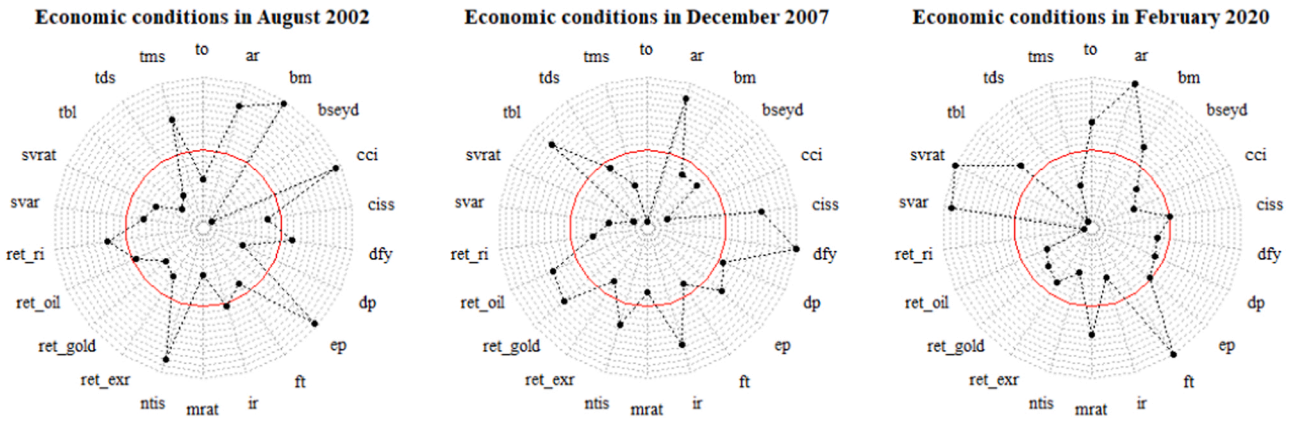

Fig. 2. Economic conditions preceding stock market crashes, Germany. This figure depicts the economic conditions that preceded the three major crash periods, namely the *dotcom bubble*, the G*lobal Financial Crisis*, and the *Covid-19 pandemic*. It visualizes the respective Kiviat charts illustratively for Germany and a single crash month during each of the three periods (September 2002, January 2008, and March 2020). For the sake of brevity, only the twenty-two one-month predictors are considered to assess the crash characteristics (the six one-year predictors are omitted). Each predictor is presented based on its detrended and scaled value. In a first step, the detrended value is computed as the residual of an ordinary least squares regression of the raw values on a time dummy, using a five-year rolling window and all information up to the respective point in time. In a second step, the detrended values are mapped into the $(-1, +1)$ interval. Positive values close to $+1$ are abnormally high, while negative values close to -1 are abnormally low. The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

5. Crash prediction models

Although multivariate crash prediction models have the potential to outperform, simple univariate approaches still dominate in both academia and industry (Berge et al., 2008; Lleo and Ziemba, 2017, 2019). Therefore, we consider univariate models as our first benchmark, which allows us to identify the incremental predictive performance of more complex multivariate approaches.

5.1. Univariate approaches

We follow Lleo and Ziemba (2017) in creating our univariate crash prediction models. The binary crash signal $CS_{t+1|t}$ is defined similarly to the binary crash indicator CI_{t+1} (see Section 4.1). It equals 1 if the respective model, incorporating all information available at the end of month t, expects a stock market crash to occur during month $t+1$, and 0 otherwise. We derive it, separately for each country and predictor, by comparing predictor variable M_t with a percentile-based threshold K_t (e.g., the 5^{th} or 95^{th} percentile) at the end of each month t. Economic theory can require M_t to exceed or fall below the threshold K_t:

$$CS_{t+1|t} = \begin{cases} 1 \text{ if } M_t - K_t > \text{or} < 0 \\ 0 \text{ otherwise} \end{cases}. \tag{2}$$

To give an example, abnormally high financial turbulence metrics ($M_t - K_t > 0$) or abnormally low book-to-market ratios ($M_t - K_t < 0$) are expected to precede stock market crashes.

Because $CS_{t+1|t} = 1$ predicts a stock market crash during month $t+1$, the economic implication is to exit the market at the end of month t, and re-enter it when $CS_{t+1|t} = 0$ again. If M_t fluctuates around K_t, this leads to excessive variability in $CS_{t+1|t}$ and frequent exits and entries, which increase transaction costs. To overcome this deficiency, Berge et al. (2008) and Lleo and Ziemba (2017) suggest combining two different percentile-based thresholds (see Appendix C, Fig. C1 for an illustration): a more restrictive threshold K_{exit}, e.g., 95%, considered for exit decisions (if the crash signal was 0 in the previous month), and a less restrictive threshold K_{entry}, e.g., 90%, considered for re-entry decisions (if the crash signal was 1 in the previous month).

To be consistent with our definition of stock market crashes (see Section 4.1), i.e., constructing univariate models that produce crash signals in five percent of the months (the expected fraction of crash indications), we focus on the simple $K_{exit} = K_{entry} = 95\%$ case.[13] A predictor signals a crash if its value exceeds the 95^{th} percentile ($M_t > K_t$) or falls below the 5^{th} percentile ($M_t < K_t$) based on the historical distribution over a ten-year rolling window.[14] Percentile-based crash signals are vulnerable to long-term trends because predictor variables exceed the thresholds more often than intended during an upward trend. Similarly, they fall below the thresholds less often during a downward trend. To overcome this problem, we differ from Lleo and Ziemba (2017) and remove the long-term trends in a preceding step. We compute the detrended values as the residuals of an ordinary least squares regression of the raw values on a time dummy, using a ten-year rolling window and all information up to the respective point in time.

5.2. Multivariate approaches

Even if some predictors possess predictive ability for subsequent stock market crashes in a univariate setting, it is unlikely, as explained in Section 4, that their predictive power is consistent across countries and constant over time. Therefore, ex ante, it is practically impossible to always select the univariate model that is optimal for a given country or at a specific point in time. Multivariate crash prediction models should outperform the *average* univariate model because they incorporate a comprehensive set of variables simultaneously. This helps diversify the cluster risks that seem inherent to models restricted to only a single predictor, i.e., that the predictive ability of this variable may vanish over time. In addition, multivariate models also include those variables that

[13] In a robustness test, we deviate from this baseline specification and consider an alternative assumption for exit and entry thresholds: $K_{exit} = 95\%$ and $K_{entry} = 90\%$. The classification results are shown illustratively for Germany in Panel A of Table B2 of Appendix B. The patterns identified for the baseline specification and their implications are robust to changes in the specification of the exit and entry thresholds.

[14] Because economic theory is ambiguous for some of the predictors used in the empirical analysis, we refrain from restricting the sign of the difference between M_t and K_t for any predictor variable ex ante.

might not yet be informative but potentially have predictive power in the future. Furthermore, most machine learning-based models consider nonlinear effects and interactions in the relationship between predictors and subsequent slumps in stock prices (McPhillips et al., 2018).

The simplest multivariate crash prediction approaches are based on logistic regressions (including models that perform variable selection/shrinkage or dimension reduction). However, due to their inability to incorporate nonlinearity and interactions (other than through *pre-determined* regression terms), more sophisticated machine learning models should outperform in terms of predictive power. Methods range from support vector machines over tree-based models to neural networks in various specifications, differing in their overall approach and complexity. More complex models may be better at modelling real-world phenomena, but complexity also raises proneness to overfitting, opaqueness, and interpretive deficiency.

From the range of different machine learning models, we first select *logistic regressions* (Ohlson, 1980). If not explicitly included as *pre-determined* terms, logistic regressions cannot capture any nonlinear or interactive effects. Therefore, we use them as our second benchmark, i.e., our linear multivariate benchmark, to identify whether nonlinear effects and interactions actually lead to incremental predictive power. To represent more sophisticated machine learning models, we select *support vector machines*, which are a well-established and popular classification method (Vapnik, 1998).[15] Like tree-based models and neural networks, they are able to incorporate nonlinearity and multi-way interactions inherently, without having to add new predictors that capture these effects in advance. Our selection is supported by studies showing that SVMs outperform relative to other classification methods, particularly in setups with small sample sizes (Huang et al., 2005; Shao and Lunetta, 2012). From a conceptual perspective, SVMs are less hierarchical than tree-based models (making them less affected by dominant predictors) and less parameterized than neural networks (making them more transparent and interpretable).

Before presenting and discussing the results of our empirical analysis, the subsections following below provide more details on machine learning classifiers. First, we describe the sample-splitting scheme used to fit the multivariate crash prediction models. Second, we provide a brief discussion of the idea behind logistic regressions and SVMs.

5.2.1. Sample splitting

While machine learning classifiers possess desirable properties, they are prone to overfitting. Therefore, we must control for the degree of model complexity by tuning the relevant hyperparameters, e.g., the vector influence and misclassification costs in SVMs. To avoid overfitting and maximize out-of-sample predictive power, hyperparameters cannot be preset, but rather must be determined adaptively from the sample data. The parameter tuning approach iteratively reduces in-sample fit by searching for a degree of model complexity that will produce reliable out-of-sample predictive performance. Gu et al. (2020) propose the time series cross-validation procedure. It splits the sample into three distinct subsamples (a training sample, a validation sample, and a test sample), which maintains the temporal ordering of the data. However, the rare occurrence of stock market crashes results in highly imbalanced datasets for training and validation, and fitting machine learning classifiers to such datasets renders them biased against the minority class, which can impair their predictive performance (Chatzis et al., 2018). To avoid such imbalances, one can use random under- or oversampling. For large datasets, *random undersampling* artificially reduces the fraction of the majority class until the sample is balanced. Small datasets require the creation of new observations with characteristics similar to those of the minority class, so *random oversampling* artificially increases the fraction of the minority class until the sample is balanced.

Since we work with a small sample of monthly data, we use the smoothed bootstrap-based random oversampling algorithm proposed by Menardi and Torelli (2014). The resulting dataset for training and validation is balanced (with similar frequencies of crash and non-crash months). Moreover, a problem is that random oversampling eliminates the time dimension of the data. We follow Chatzis et al. (2018) and opt for the k-fold cross-validation approach. It randomly splits the *balanced* dataset into k parts ($k-1$ parts for training and the remaining part for validation) and repeats this procedure k times, such that each part is used for validation once.[16] While each training sample is used to estimate the model for multiple parameter specifications, each validation sample is used to compute the validation accuracy (Acc_{val}).[17] The optimal specification of hyperparameters maximizes the *average* validation accuracy: $\overline{Acc}_{val} = \frac{1}{k}\sum_{j=1}^{k} Acc_{val,j}$.[18] Finally, the test sample, which is used for neither model estimation nor parameter tuning, is truly out-of-sample and appropriate for evaluating a model's predictive power.

In addition, we adopt the ensemble approach proposed by Dietterich (2000) and Bluwstein et al. (2020). Using five independent seeds s, we first apply the random oversampling algorithm five times at each re-estimation date. Based on each of the five distinct balanced training and validation samples, we fit one independent model (either a logistic regression or an SVM) to compute the crash signals.[19] Incorporating all information available at the end of month t, each model first estimates the likelihood $p_{t+1|t,s}$ that a stock market crash will occur during month $t+1$ and then generates a crash signal that equals 1 if the likelihood estimate exceeds 0.5, and 0 otherwise:

$$CS_{t+1|t,s} = \begin{cases} 1 \text{ if } p_{t+1|t,s} > 0.5 \\ 0 \text{ otherwise} \end{cases}. \tag{3}$$

However, instead of generating seed-level crash signals, we average the different predictions $p_{t+1|t,s}$ into a single ensemble prediction: $\bar{p}_{t+1|t} = \frac{1}{5}\sum_{s=1}^{5} p_{t+1|t,s}$. This is the predicted stock market crash probability, which we denote as $\bar{p}_{crash}$. Because the stochastic nature of random oversampling leads to different balanced datasets and likelihood estimates for each seed, averaging predictions across the different seeds reduces noise and increases the signal-to-noise ratio. For the ensemble prediction, we follow the same logic applied to generate the crash signals within each seed, setting the crash signal equal to 1 if the *average* likelihood $\bar{p}_{t+1|t}$(across the five seeds) exceeds 0.5, and 0 otherwise:

$$CS_{t+1|t} = \begin{cases} 1 \text{ if } \bar{p}_{t+1|t} > 0.5 \\ 0 \text{ otherwise} \end{cases}. \tag{4}$$

Finally, it is important to note that the sample-splitting scheme must periodically include more recent data (see West, 2006, for an overview). In general, both the "rolling window" and the "recursive window"

[15] In a robustness test, we use other machine learning classifiers such as two tree-based models (random forests and gradient boosted regression trees) and a neural network. For the sake of brevity, the results for Germany are shown in Panel B of Table B2 of Appendix B. The results for these alternative machine learning classifiers are similar to those for the SVM-based model.

[16] Using random oversampling, we set the number of monthly observations in the balanced dataset to $obs_{blncd} = 500$, and the number of folds to $k = 5$. This implies 5 folds of 100 observations used for training and validation.

[17] Logistic regressions do not require parameter tuning (based on the validation samples) and are estimated from the combined training and validation samples.

[18] We use the overall accuracy of the binary classification as our loss function. Because it is computed based on balanced training and validation samples, it is suitable despite the imbalances in the underlying sample.

[19] Seeds are numbers used to initialize random processes, which ensures different but reproducible predictions.

approach are applicable. We choose the latter approach, which progressively increases the underlying dataset intended for training and validation, and always incorporates the entire history of data, ensuring a sufficient number of crashes during the fitting process.[20] We recursively refit the models each month by adding one month to the imbalanced dataset at each re-estimation date.[21]

5.2.2. Machine learning classifiers

The idea behind a crash prediction model based on *logistic regressions* (*logit*) is running pooled ordinary least squares regressions of the log-odds (i.e., the logarithm of the odds, denoted as "logit") that a stock market crash will occur during the next month ($CI_{t+1} = 1$) on the set of twenty-eight predictors z_t:[22]

$$logit(CI_{t+1} = 1) = \alpha_t + \beta_t' z_t + \varepsilon_{t+1}. \quad (5)$$

In contrast, the idea behind a crash prediction model based on *support vector machines* (*svm*) is to search for hyperplanes that territorially divide a multidimensional vector space into groups of vectors that belong to the same class (see Appendix C, Fig. C2 for a two-dimensional, two-class illustration).[23] Each potential hyperplane is located in an area where vectors of two different classes are close together. To increase computational speed, SVMs do not always use all vectors from the vector space. Rather, they focus on those in the immediate neighborhood of the potential hyperplane, so-called "support vectors". The algorithm then specifies the optimal hyperplane by aiming to 1) maximize the distance of correctly classified support vectors from the hyperplane and 2) minimize the number of misclassified support vectors.

In theory, the algorithm can avoid any misclassification if there are no restrictions on the shape of the hyperplanes (but requiring indefinite computational power and time). Since this likely leads to overfitting (regardless of obvious computational limitations), SVMs must be strongly regularized. We follow Drobetz and Otto (2021) and use a radial basis function (RBF) kernel for a proper nonlinear transformation of the vector space. We simultaneously apply two other common types of regularization. First, we constrain the influence of any single vector, i.e., we restrict the space within which it can serve as a support vector (using a vector influence parameter γ). A smaller vector influence avoids enabling vectors to serve as supports for overly distant hyperplanes. Second, we set the permitted number of misclassified support vectors to a positive value, i.e., we allow for a certain number of misclassifications (using a misclassification cost parameter c). Smaller misclassification costs ignore more of the misclassified support vectors, while continuing to fit optimal hyperplanes.[24]

6. Empirical results

Having introduced the characteristics of stock market crashes and the crash prediction models, we now compare their predictive performance in out-of-sample tests from both a statistical and an economic perspective. In addition, we conduct in-sample tests to investigate the characteristics and functioning scheme of our machine learning-based approaches.

6.1. Out-of-sample tests

We begin with contrasting the models' ability to forecast stock market crashes out-of-sample in terms of classification performance, and proceed with comparing the performance of market timing and switching strategies. Because the relationship between statistical measures and economic value-added may be only weak (Leitch and Tanner, 1991; Cenesizoglu and Timmermann, 2012), we analyze both dimensions of predictive performance.

6.1.1. Statistical predictive performance

Crash prediction models are textbook examples of binary classifiers. Potential outcomes of binary classifications are true positives (TPs) if the predicted and the realized class equal 1, true negatives (TNs) if the predicted and the realized class equal 0, false positives (FPs) if the predicted class equals 1 but the realized class equals 0, and false negatives (FNs) if the predicted class equals 0 but the realized class equals 1. TPs are crash months correctly classified as crash months, and FPs are non-crash months incorrectly classified as crash months. Table C1 of Appendix C details the four possible cases. As visualized in Panel A, the classification performance is usually evaluated using confusion matrices, together with a broad set of classification measures, in particular, *accuracy*, *precision*, *recall*, and *F1* (the harmonic mean of precision and recall) measures. To give an example, Panel B provides a visualization of a binary classification and the four corresponding measures.

Because stock market crashes are extreme events and, by definition, rare, there are only a few crash months during the sample period. This results in highly imbalanced out-of-sample datasets, and thus the most common classification measure, overall accuracy, can be misleading (Luque et al., 2019). A better way to measure a model's predictive performance is the conditional probability $P(CI = 1|CS = 1)$, i.e., the probability that stock market crashes actually occur when they are expected to occur. Following Lleo and Ziemba (2017), Appendix A derives the maximum likelihood estimator $\hat{p}$ that represents a model's overall precision, i.e., the number of true positive classifications divided by the number of distinct crash signals. Moreover, we test whether the models' classification performance is significantly different from that of a random classifier (no-information rate). To this end, Appendix A derives the likelihood ratio test statistic (Y-statistic) for the null hypothesis that a model's conditional probability is equal to the no-information rate, for which we report the empirical p-statistics, corrected for a potential small sample bias stemming from the low number of crash signals.

Panels A and B of Table 2 report the classification metrics for the univariate and multivariate models illustratively for Germany. For the sake of brevity, the classification results for the remaining sample

[20] As illustrated in Section 4.2, stock market crashes are preceded by different underlying economic conditions, i.e., they occur for different reasons. It is thus unclear whether a recursive-window approach is necessarily preferred over a rolling-window approach, which holds the dataset intended for training and validation constant and ignores potentially outdated information. However, in our empirical setting, recursive windows are the only way to ensure a sufficient number of crashes during the fitting process, which is why we opt for this approach.

[21] To allow for a comparison with the univariate crash prediction models, the first multivariate crash signals are also computed for January 2000, leaving 120 months from January 1990 to December 1999 for initial fitting.

[22] In a robustness test (unreported), we observe that the statistical predictive performance of multivariate logistic regressions that incorporate all twenty-eight predictors simultaneously is superior to the average statistical predictive performance of twenty-eight univariate logistic regressions that incorporate each predictor variable separately.

[23] In our analysis, each vector (observation) is defined by the twenty-eight predictors and assigned to either the crash or non-crash month class.

[24] The tuning parameters for the SVMs are the vector influence, which we set to $\gamma \in (0.001, 1)$, and the misclassification costs, which we set to $c \in (0.001, 1)$.

Table 2
Statistical predictive performance, Germany.

Model specification		Classification metrics												
Model	Sign	n_{preds}	TP	FP	TN	FN	Acc. [%]	Prec. [%]	Rec. [%]	F1 [%]	*Y*-stat	*p*-stat [%]	CP1	CP2	CP3
Germany ($n_{periods} = 252$, $n_{crashes} = 16$)															
Panel A: Univariate crash models ($K_{exit} = 95\%$, $K_{entry} = 95\%$)															
ep	+	17	5	12	224	11	91	29	31	30	8.55	0.04	✔	×	×
ft	+	9	3	6	230	13	92	33	19	24	5.87	0.16	×	✔	✔
ret_exr	–	10	3	7	229	13	92	30	19	23	5.24	0.28	×	✔	×
bm	+	26	5	21	215	11	87	19	31	24	4.87	0.53	✔	×	×
mrat	–	19	4	15	221	12	89	21	25	23	4.47	0.56	✔	✔	×
ret_ri	–	13	3	10	226	13	91	23	19	21	3.81	0.68	✔	✔	✔
ret_ri_ann	–	15	3	12	224	13	90	20	19	19	3.10	1.21	✔	✔	×
bseyd	–	15	3	12	224	13	90	20	19	19	3.10	1.21	✔	×	×
tds	+	17	3	14	222	13	89	18	19	18	2.53	2.06	✔	✔	×
ret_exr_ann	+	17	3	14	222	13	89	18	19	18	2.53	2.06	✔	×	×
ret_oil_ann	–	17	3	14	222	13	89	18	19	18	2.53	2.06	✔	✔	×
dp	–	27	4	23	213	12	86	15	25	19	2.42	19.56	✔	×	×
svar_ann	+	18	3	15	221	13	89	17	19	18	2.29	32.32	×	✔	✔
ciss	+	18	3	15	221	13	89	17	19	18	2.29	32.32	×	✔	×
dfy	+	18	3	15	221	13	89	17	19	18	2.29	32.32	×	✔	×
cci	–	10	2	8	228	14	91	20	13	15	2.07	2.25	✔	✔	×
svar	+	11	2	9	227	14	91	18	13	15	1.78	2.87	×	✔	×
svrat	+	11	2	9	227	14	91	18	13	15	1.78	2.87	×	✔	✔
to	–	7	1	6	230	15	92	14	6	9	0.56	69.85	✔	×	×
ntis	–	27	1	26	210	15	84	4	6	5	0.37	41.21	×	×	✔
ret_gold_ann	–	22	2	20	216	14	87	9	13	11	0.25	38.82	×	✔	×
tms	–	10	1	9	227	15	90	10	6	8	0.19	64.39	✔	×	×
tbl	+	11	1	10	226	15	90	9	6	7	0.12	63.04	✔	×	×
ar	+	18	1	17	219	15	87	6	6	6	0.02	61.55	×	×	✔
ret_gold	–	14	1	13	223	15	89	7	6	7	0.01	61.80	×	✔	×
ret_oil	–	15	1	14	222	15	88	7	6	6	0.00	61.15	×	✔	×
ir	+	17	0	17	219	16	87	0	0	0			×	×	×
ir_ann	+	17	0	17	219	16	87	0	0	0			×	×	×
Panel B: Multivariate crash models															
logit		34	5	29	207	11	84	15	31	20	2.98	12.70	✔	✔	✔
svm		**29**	**7**	**22**	**214**	**9**	**88**	**24**	**44**	**31**	**9.43**	**0.03**	✔	✔	✔

This table reports the classification metrics for different crash prediction models illustratively for Germany during the January 2000–December 2020 out-of-sample period. The results for the remaining sample countries (France, Italy, Spain, and the Netherlands) are shown in Table B1 of Appendix B. The classification metrics are based on the comparison of $CS_{t+1|t}$ and CI_t. CI_{t+1} is a binary crash indicator, which equals 1 when a substantial stock market downturn occurs during month $t + 1$, and 0 otherwise. $CS_{t+1|t}$ is a binary crash signal, which equals 1 if the respective model, incorporating all information available at the end of month t, expects a crash to occur during month $t + 1$, and 0 otherwise. Panel A presents the metrics for the univariate crash prediction models introduced in Section 5.1. In addition to the number of crash signals (n_{preds}), the numbers of true/false positives ($\#TP/\#FP$) and true/false negatives ($\#TN/\#FN$) are reported, together with accuracy (Acc.), precision (Prec.), recall (Rec.), and F1 measures. The likelihood ratio test statistic (*Y*-stat) testing the null hypothesis that a crash prediction model's conditional probability is equal to the no-information rate is added, together with the empirical *p*-statistic (*p*-stat) and the distribution of true positives across the three major crash periods (CP1/CP2/CP3). The last three columns indicate whether the respective model is able to correctly forecast at least one stock market crash within each of the three subperiods of the sample (surrounding the three major crash periods), i.e., January 2000–December 2007 (CP1), January 2008–December 2014 (CP2), and January 2015–December 2020 (CP3). Panel B presents the metrics for the multivariate crash prediction models based on logistic regressions and support vector machines (logit and svm; introduced in Section 5.2). The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

countries are shown in Table B1 of Appendix B.[25] Focusing on the univariate models, differences in likelihood ratio test statistics (on average, across all predictors within a given country) suggest that the overall ability to forecast future stock market crashes differs substantially across countries. The *average* likelihood ratio test statistic is highest for Germany (2.25) and lowest for the Netherlands (1.37). Moreover, within each country, the statistical predictive performance varies notably across predictors. We also observe that, in some cases, similar variables reflect the best and worst classification metrics in the different sample countries, implying that forecast abilities are consistent across countries (at least to some extent). For example, valuation metrics such as the earnings-to-price ratio (*ep*) and the dividend-to-price ratio (*dp*) tend to possess high predictive ability for subsequent stock market crashes, while the returns on gold or oil markets (*ret_gold* and *ret_oil*) or the yields on ten-year government bonds (*ir* and *ir_ann*) are less informative. However, given empirical *p*-statistics mostly above 5%, only a few variables deliver univariate models with conditional probabilities significantly larger than the no-information rate. In many sample countries, some variables are not even capable to create a single crash signal ($n_{preds} = 0$) or fail to correctly forecast a single stock market crash

[25] To understand why the tables can have empty cells, it is important to note that there are three potential scenarios for crash prediction models. First, a model fails to create a single crash signal ($n_{preds} = 0$). In this case, neither classification outcomes (such as the number of true positives) nor classification metrics (such as accuracy) can be calculated. In addition, the likelihood ratio test statistic is not applicable. Second, a model is able to create at least one crash signal ($n_{preds} \geq 1$), but fails to produce at least one true positive classification ($\#TP = 0$). In this case, both classification outcomes and classification metrics can be calculated. The likelihood ratio test statistic, however, is still not applicable. Third, a model is able to create at least one crash signal ($n_{preds} \geq 1$) and, in addition, to produce at least one true positive classification ($\#TP \geq 1$). This is the most common case, in which all cells are filled with the respective numbers.

during the sample period, i.e., they produce zero true positive classifications ($\#TP = 0$).

There are two important caveats: First, in multiple cases, predictors that generally perform well across countries perform badly in at least a single country. For example, the univariate model based on the earnings-to-price ratio (*ep*), which performs very well in Germany, France, Italy and Spain, fails to beat the no-information rate in the Netherlands. Second, nearly all univariate models fail to correctly forecast crashes within each of the three subperiods of the sample (surrounding the three major crash periods).[26] We conclude that, ex ante, it is basically impossible to always select the univariate model that is optimal for a given country or at a specific point in time.

Next, focusing on multivariate models, we find that both the *svm* and the *logit* model yield notably higher likelihood ratio tests statistics than the average univariate benchmark for all sample countries (expect for Spain in the case of logistic regressions). This documents that the multivariate models' ability to incorporate multiple predictor variables simultaneously translates into superior statistical predictive performance relative to their univariate counterparts. The likelihood ratio test statistics also indicate that the classification performance of the *svm* model relative to the *logit* model is slightly lower for Italy, but higher for France and the Netherlands.[27] In addition, for Germany and Spain, *svm* models substantially outperform logistic regressions in terms of sizably higher likelihood ratio test statistics. This suggests that the *svm* model's ability to capture nonlinear and interactive effects tends to deliver incremental predictive performance compared to the *logit* model.[28]

Moreover, empirical *p*-statistics below 5% indicate that the *svm* model's conditional probability differs significantly from the no-information rate for all sample countries (except for Spain). For these countries, the *svm* model is also able to correctly forecast crashes within each of the three subperiods of the sample (surrounding the three major crash periods). We conclude that, because the correct crash signals are more evenly distributed over the sample period, the classification performance of the *svm* model is more likely to persist than that of univariate models. Finally, consistent with these results from the likelihood ratio test statistics, we find that the *svm* model also beats the average univariate benchmark as well as a multivariate logistic regression model for most sample countries when comparing the crash prediction models based on their F1 measures instead of their likelihood ratio test statistics (see Table 2 and Appendix B, Table B1).[29]

To further substantiate our arguments, Fig. 3 illustrates the classification performance for the German stock market, taking the *svm* model (Panel A) and the univariate model with the highest likelihood ratio test statistic, the *ep* model based on the earnings-to-price ratio (Panel B), as examples. Both figures depict monthly stock market returns; the classification results are flagged as follows: True positives (TPs) are marked with a green filled circle, true negatives (TNs) with a green unfilled circle, false positives (FPs) with a yellow filled circle, and false negatives (FNs) with a red filled circle. The *svm* model generates crash signals and delivers TPs around all three major crash periods. In sharp contrast, the *ep* model creates nearly all crash signals during the dotcom bubble, and all its TPs are clustered during this time period as well. While the classification performance over the full sample period is only slightly higher for the *svm* model compared to *ep* model (as indicated by similar likelihood ratio test statistics of 9.43 and 8.55, respectively), it is important to recognize that the earnings-to-price ratio lost its predictive power shortly after the dotcom bubble. Being a multivariate model, SVMs are much less exposed to such cluster risks that seem inherent to models restricted to only a single predictor, i.e., that the predictive ability of this variable may vanish over time.

6.1.2. Investment portfolio performance

Next, we assess the value-added of multivariate crash prediction models to active investors under realistic trading assumptions. The cumulative performance of an investment portfolio depends on the trade-off between the ability to correctly classify crash and non-crash months (TPs and TNs, respectively) and the ability to avoid falsely misstating non-crash and crash months (FPs and FNs, respectively). We consider two common benchmarks, which we contrast with investment strategies based on crash signals provided by the *logit* and *svm* models.[30]

We use the stock market excess return as our first benchmark and compare it with a market timing strategy that buys or holds the stock market to earn the benchmark return when the crash signal equals 0, and leaves or stays out of the stock market otherwise (investing in cash, which results in a zero excess return). Statistical predictive performance only translates into value-added to active investors relative to the stock market benchmark if the benefits from missing negative stock market returns outweigh the costs of missing positive stock market returns (accounting for the asymmetric nature of cumulative performance).

A 50/50 balanced stock-bond market portfolio, averaging the stock and bond market excess returns, serves as our second benchmark, which we compare with a market switching strategy. This strategy earns the stock market excess return when the crash signal equals 0, and the bond market excess return otherwise. Since stock and bond markets are

[26] The last three columns in Table 2 and Tables B1 and B2 of Appendix B indicate whether the respective model is able to correctly forecast at least one stock market crash within each of the three subperiods of the sample (surrounding the three major crash periods), i.e., January 2000–December 2007 (CP1), January 2008–December 2014 (CP2), and January 2015–December 2020 (CP3).

[27] The ranking of the svm model relative to the logit model and their univariate counterparts may falsely understate the true predictive performance for active investors. For example, in Italy, a stock market crash occurred in October 2008, but the average crash probability across the different seeds for the next month computed in September 2008 (45.77%) remains marginally below the 0.5 threshold, which is necessary for the SVM-based model to signal a stock market crash. In practice, many of those investors would still consider this a crash signal provided that the dispersion in estimated crash probabilities across the different seeds (from 38.90% to 51.60%) is sufficiently low (high-precision forecast). If an investor interpreted this as a crash signal, the classification performance of the *svm* model, resulting in a likelihood ratio test statistic of 5.74, would be higher than that of the *logit* model and close to that of the best-performing univariate models.

[28] The svm model's ability to forecast future stock market crashes differs across countries. Overall, the predictive ability is the highest for Germany and the lowest for Spain (as indicated by a likelihood ratio test statistic of 9.43 and 1.69, respectively; see Table 2 and Appendix B, Table B1). We observe a similar rank order across countries for the svm model's predictive ability (the highest for the Netherlands and the lowest for Spain) when fitting the SVMs on the full dataset originally used for out-of-sample testing (covering all 252 months). Based on these in-sample results (unreported), we conclude that stock market crashes follow more systematic and predictable patterns in the Netherlands. In contrast, crashes seem to be harder to predict in Spain, which may be explained by the nature of its stock market and economy. The Spanish stock market is more volatile, suffering from multiple high-volatility periods, which are more difficult to predict. During high-volatility periods, low stock market returns may indicate crashes without substantial changes in the underlying economic conditions, which can impair the svm model's predictive performance. Moreover, some predictor variables are based on German or U.S. data, but also serve as crash risk proxies for the other countries in our sample. These predictors may be less informative in those countries that face high idiosyncratic crash risk due to country-specific economic conditions such as overindebtedness.

[29] The results are qualitatively similar when comparing the crash prediction models using Alessi and Detken's (2011) concept of usefulness instead of their likelihood ratio test statistics (unreported).

[30] We take the total return indices provided by Refinitiv to calculate bond and stock market returns, and the three-month FIBOR or EURIBOR rate, whichever is available, scaled to the one-month horizon, as the risk-free rate to calculate excess returns.

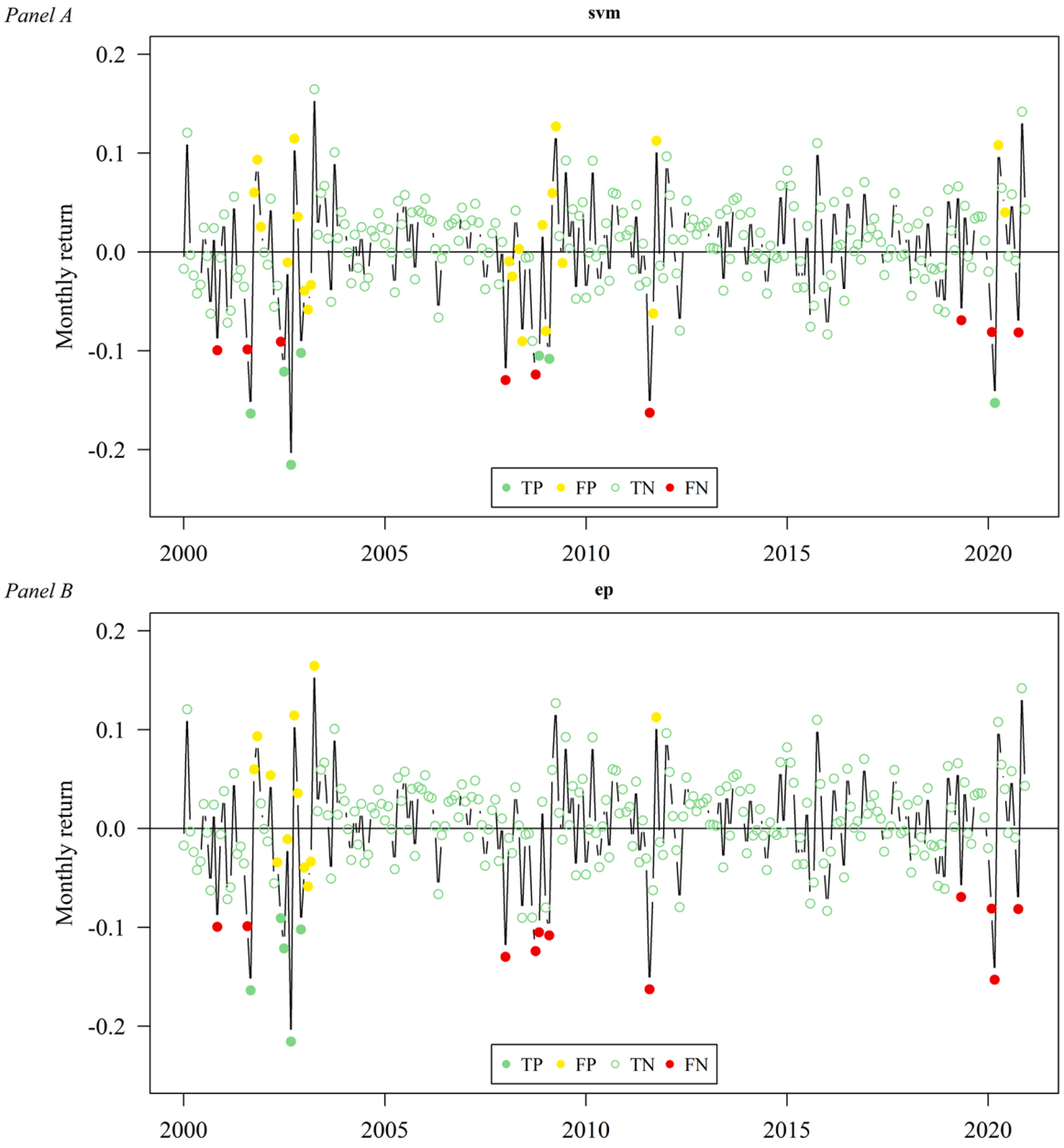

Fig. 3. Statistical predictive performance, Germany. This figure visualizes the classification performance of different crash prediction models during the January 2000–December 2020 out-of-sample period. The classification performance is based on the comparison of $CS_{t+1|t}$ and CI_t. CI_{t+1} is a binary *crash indicator*, which equals 1 when a substantial stock market downturn occurs during month $t + 1$, and 0 otherwise. $CS_{t+1|t}$ is a binary *crash signal*, which equals 1 if the respective model, incorporating all information available at the end of month t, expects a crash to occur during month $t + 1$, and 0 otherwise. The monthly stock market returns are plotted; the classification results are flagged as follows: True positives (TPs) are marked with a green filled circle, true negatives (TNs) with a green unfilled circle, false positives (FPs) with a yellow filled circle, and false negatives (FNs) with a red filled circle. The charts are presented for the multivariate crash prediction model based on support vector machines (*svm;* introduced in Section 5.2) in Panel A and the best-performing univariate model based on the earnings-to-price ratio (*ep;* introduced in Section 5.1) in Panel B illustratively for Germany. The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

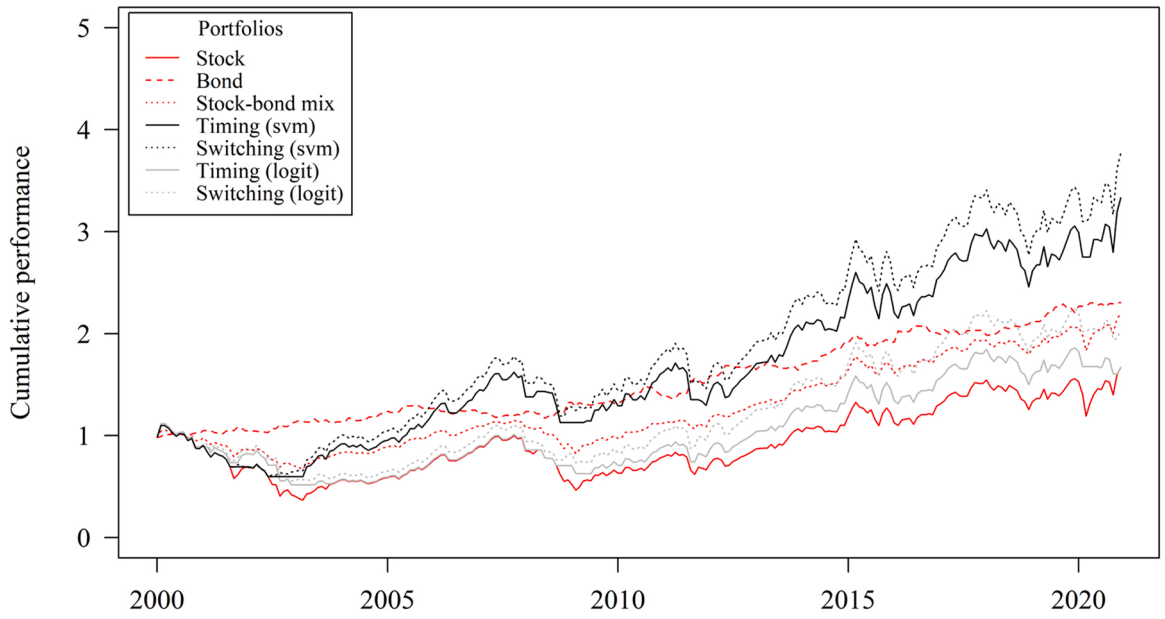

Fig. 4. Investment portfolio performance, Germany. This figure visualizes the cumulative performance of machine learning-based investment portfolios during the January 2000–December 2020 out-of-sample period. The two investment strategies under investigation are based on $CS_{t+1|t}$. $CS_{t+1|t}$ is a binary *crash signal*, which equals 1 if the respective model, incorporating all information available at the end of month t, expects a crash to occur during month $t + 1$, and 0 otherwise. The market timing strategy buys or holds the stock market to earn the benchmark return when the crash signal equals 0, and leaves or stays out of the stock market otherwise (investing in cash, which results in a zero excess return). The market switching strategy earns the stock market excess return when the crash signal equals 0, and the bond market excess return otherwise. The cumulative performance of the machine learning-based investment portfolios is compared to three benchmarks, i.e., a stock market portfolio (earning the stock market excess returns), a bond market portfolio (earning the bond market excess returns), and a 50/50 balanced stock-bond market portfolio (averaging the stock and bond market excess returns). The portfolio values are scaled to €1 at the beginning of January 2000. The charts are presented for the multivariate crash prediction models based on logistic regressions and support vector machines (*logit* and *svm;* introduced in Section 5.2) illustratively for Germany, and net of transaction costs (a conservative lump-sum discount of 10 bps per exit/re-entry). The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

negatively correlated during the sample period (unreported), both the average benefits from true positives and the average costs of false positives increase.

We compute return and risk metrics for the stock, bond, and stock-bond market portfolios as well as the four machine learning-based portfolios (i.e., one market timing and one market switching portfolio for the *logit* and *svm* models, respectively). We also take into account transaction costs. Both the benchmark and forecast strategies can be realized in a simply way via buying, holding, and selling exchange-traded funds (ETFs) that replicate the respective markets. To be conservative, we assume that the market ETFs are free of transaction costs, despite the need for monthly rebalancing. For any buy or sell transaction related to market timing or switching, we refer to Borkovec and Serbin (2013) and Angel et al. (2016), choosing a conservative lump-sum discount of 10 bps on the portfolio's monthly excess return (which is at least three times the size of the historical bid–ask spreads reported in these studies).

Using the German stock market as an illustration, Fig. 4 depicts the cumulative performance (net of transaction costs) of investments of €1 in both benchmarks as well as the market timing and switching strategies at the beginning of January 2000. Table 3 presents the return and risk metrics (net of transaction costs) of the investment portfolios for each of the five sample countries. The terminal value is reported at the end of December 2020, together with the annualized excess return, annualized volatility, maximum drawdown, Sharpe ratio, and information ratio (relative to the market portfolio).

Focusing again on the results for Germany, the market timing and switching strategies for the *svm* model yield terminal values of 3.33 and 3.77, respectively, which notably exceed the corresponding numbers for the *logit* model (1.67 and 2.03, respectively) . The average excess return for the SVM-based portfolios is higher compared to the stock and stock-bond portfolios (5.90% and 6.52% vs. 2.47% and 3.82%), while volatility is within the benchmark range (15.01% and 15.25% vs. 18.56% and 9.14%). Bond markets generate positive and stable excess returns, and are negatively correlated with stock markets during our sample period (unreported). In contrast, stock markets faced several turbulent times, with three major crash periods (see Fig. 1). As a result, market switching outperforms market timing. However, a timing strategy reflects the multivariate models' ability to predict crashes more stringently because its cumulative performance depends only on stock market excess returns. Compared to the stock market, higher average excess returns and lower volatility for the *svm* model triple the Sharpe ratio (0.40 vs. 0.13), and translate into a positive information ratio (0.32). The maximum drawdown is reduced by roughly one-third (45.83% vs. 66.60%).[31]

[31] The results are qualitatively similar for the remaining sample countries (except for Spain). The SVM-based market timing and switching strategies outperform both benchmarks (the stock market and the 50/50 balanced stockbond market portfolio), which translates into positive information ratios. Following the crash signals provided by the SVMs also substantially reduces both volatility and maximum drawdown. In a robustness test (unreported), we further investigate the performance of a cross-country market timing strategy. This active strategy reallocates the funds pulled out of those markets for which a crash is predicted to those markets for which no crash is signaled (with equal weights) on a monthly basis. It delivers notably improved return and risk metrics compared to a naïve strategy that assigns equal weights to each country's stock market.

Table 3
Investment portfolio performance.

	Benchmark portfolios			Machine learning-based portfolios			
				logit		svm	
	Stock	Bond	Stock-bond mix	Timing	Switching	Timing	Switching
Germany (n_{exits} = 9, $n_{re-entries}$ = 9)							
Terminal value [€]	1.67	2.30	2.20	1.67	2.03	3.33	3.77
Excess return annualized [%]	2.47	4.05	3.82	2.48	3.43	5.90	6.52
Std. annualized [%]	18.56	5.78	9.14	15.27	15.48	15.01	15.25
Maximum drawdown [%]	66.60	9.84	36.56	54.00	50.42	45.83	46.18
Sharpe ratio	0.13	0.70	0.42	0.16	0.22	0.40	0.43
Information ratio		0.08	0.13	0.00	0.09	0.32	0.35
France (n_{exits} = 8, $n_{re-entries}$ = 8)							
Terminal value [€]	1.94	2.48	2.41	2.49	2.65	3.46	3.94
Excess return annualized [%]	3.21	4.42	4.29	4.44	4.74	6.09	6.75
Std. annualized [%]	17.32	5.94	8.82	15.98	16.04	15.09	15.31
Maximum drawdown [%]	58.28	11.00	31.20	58.33	55.63	52.83	47.05
Sharpe ratio	0.19	0.75	0.49	0.28	0.30	0.41	0.44
Information ratio		0.06	0.11	0.19	0.22	0.35	0.39
Italy (n_{exits} = 11, $n_{re-entries}$ = 11)							
Terminal value [€]	1.05	2.92	1.94	1.62	1.64	1.98	2.02
Excess return annualized [%]	0.25	5.23	3.20	2.31	2.40	3.29	3.40
Std. annualized [%]	19.63	7.96	11.40	17.33	17.62	17.12	17.39
Maximum drawdown [%]	62.60	17.28	38.13	56.00	56.70	50.69	50.95
Sharpe ratio	0.01	0.66	0.28	0.13	0.14	0.19	0.20
Information ratio		0.26	0.30	0.23	0.24	0.32	0.33
Spain (n_{exits} = 13, $n_{re-entries}$ = 13)							
Terminal value [€]	1.52	2.91	2.31	1.34	1.92	1.57	1.67
Excess return annualized [%]	2.03	5.22	4.08	1.39	3.16	2.18	2.47
Std. annualized [%]	19.11	7.70	11.06	17.86	18.03	16.82	17.06
Maximum drawdown [%]	57.95	11.70	32.34	53.23	46.72	52.53	51.99
Sharpe ratio	0.11	0.68	0.37	0.08	0.18	0.13	0.15
Information ratio		0.17	0.22	-0.09	0.16	0.02	0.05
Netherlands (n_{exits} = 10, $n_{re-entries}$ = 10)							
Terminal value [€]	1.64	2.48	2.25	2.95	3.12	3.70	4.32
Excess return annualized [%]	2.38	4.43	3.93	5.29	5.57	6.42	7.21
Std. annualized [%]	18.01	5.83	8.99	15.49	15.53	14.45	14.69
Maximum drawdown [%]	64.77	9.91	36.39	46.82	47.40	40.79	33.67
Sharpe ratio	0.13	0.76	0.44	0.34	0.36	0.45	0.49
Information ratio		0.10	0.16	0.32	0.34	0.38	0.43

This table reports the return and risk characteristics of machine learning-based investment portfolios during the January 2000–December 2020 out-of-sample period. The two investment strategies under investigation are based on $CS_{t+1|t}$. $CS_{t+1|t}$ is a binary crash signal, which equals 1 if the respective model, incorporating all information available at the end of month t, expects a crash to occur during month $t+1$, and 0 otherwise. The market timing strategy buys or holds the stock market to earn the benchmark return when the crash signal equals 0, and leaves or stays out of the stock market otherwise (investing in cash, which results in a zero excess return). The market switching strategy earns the stock market excess return when the crash signal equals 0, and the bond market excess return otherwise. The cumulative performance of the machine learning-based investment portfolios is compared to three benchmarks, i.e., a stock market portfolio (earning the stock market excess returns), a bond market portfolio (earning the bond market excess returns), and a 50/50 balanced stock-bond market portfolio (averaging the stock and bond market excess returns). The portfolio values are scaled to €1 at the beginning of January 2000. The numbers are presented for the multivariate crash prediction models based on logistic regressions and support vector machines (logit and svm, introduced in Section 5.2) for each of the five sample countries, and net of transaction costs (a conservative lump-sum discount of 10 bps per exit/re-entry). The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

6.2. In-sample tests

We next conduct in-sample tests to investigate the characteristics and functioning scheme of the multivariate crash prediction models. In particular, we inspect changes in the inherent model complexity over time, decompose predictions into the contributions of individual variables using relative variable importance metrics, and explore patterns of nonlinear and interactive effects in the relationship between predictor variables and estimated crash probabilities.

6.2.1. Model complexity

Since we re-estimate the *svm* and *logit* models on a monthly basis, it is interesting to gauge whether model complexity changes over time or remains stable. We define model complexity as the difficulty of forecasting subsequent stock market crashes, i.e., the dispersion in estimated likelihoods across the different seeds. In particular, we measure model complexity (*mc*) as the spread between the second-highest and second-lowest likelihoods within the ensemble prediction. Since difficulties in creating reliable crash signals (as indicated by larger spreads) may be attributable to extreme economic conditions such as high-volatility periods, we relate the *mc* metrics to the current-month stock market variance (*svar*).

Illustratively for Germany, Fig. 5 depicts the degree of model complexity at each re-estimation date and its association with stock market volatility by plotting *mc_svm*, *mc_logit*, and *svar* over time. As in Gu et al. (2020) and Drobetz and Otto (2021), model complexity varies substantially over time, and both SVMs and logistic regressions exhibit periods of high-precision and low-precision forecasts.[32] We find

[32] Precision, taken in isolation, does not reveal information about the SVMs' predictive ability in general or accuracy in particular. Precision only measures whether there are large or small variations in single predictions across the different seeds, which points towards uncertain economic conditions. Accuracy, in turn, indicates whether the ensemble prediction is, on average, close to its realization

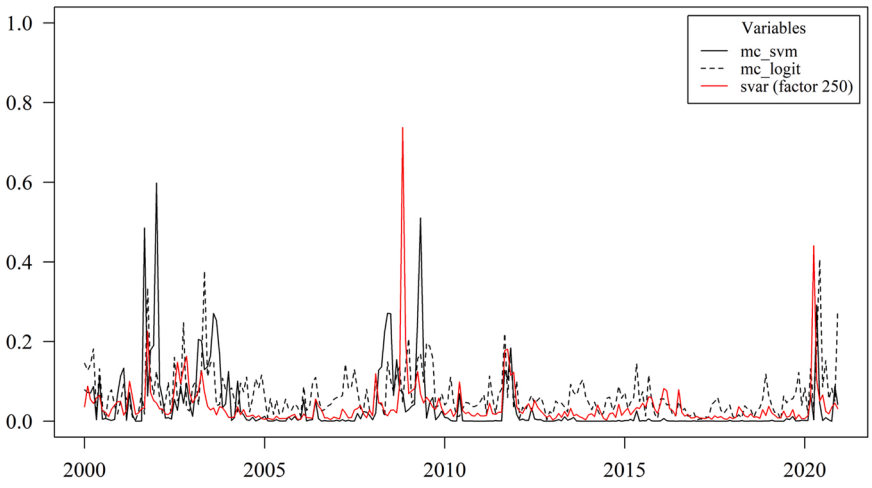

Fig. 5. Model complexity over time, Germany. This figure visualizes the degree of model complexity for different crash prediction models and the one-month stock market variance (*svar*) at each re-estimation date during the January 2000–December 2020 out-of-sample period. Model complexity is defined as the difficulty of forecasting subsequent stock market crashes, i.e., the dispersion in estimated likelihoods across the different seeds. In particular, model complexity (*mc*) is measured as the spread between the second-highest and second-lowest likelihoods within the ensemble prediction. The charts are presented for the multivariate crash prediction models based on logistic regressions and support vector machines (*logit* and *svm;* introduced in Section 5.2) illustratively for Germany. The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

synchronicity in the *mc* and *svar* metrics (with *mc_logit* being, on average, slightly higher than *mc_svm*), indicating that high-volatility periods are more difficult to predict.[33]

6.2.2. Variable importance

Next, since the degree of model complexity is time-varying, it is instructive to explore whether each predictor's contribution to the overall forecasting ability of the two multivariate models also changes over time. To this end, we calculate the variable importance matrix based on a two-step approach, separately for each re-estimation date: First, we compute the absolute variable importance as the decrease in likelihood ratio test statistic from setting all values of a given predictor to its uninformative median value within the balanced datasets.[34] Second, we normalize the absolute variable importance measures to sum to one, signaling the relative contribution of each variable to the respective model.

Panel A of Fig. 6 depicts the time series averages of relative variable importance measures for the *svm* and *logit* models illustratively for Germany. We find that, despite some differences, both models classify similar predictors as informative. The most influential predictors are based on exchange rate trends (*ret_exr*) as well as returns on stock, oil, and gold markets (*ret_stock*, *ret_oil*, and *ret_gold*). Variables reflecting current stock market risk, e.g., the financial turbulence metric (*ft*) or the current-month stock market variance (*svar*), are also highly important. In contrast, information from bond markets (*ir*) seem to be less relevant.

Since the overall relative variable importance measures only mirror a predictor's mean contribution to a model's predictive performance, we also investigate the relative variable importance metrics over time. Volatile metrics indicate that all covariates in the predictor set are important. In contrast, stable figures mean we should remove uninformative predictors permanently, as they may decrease a model's signal-to-noise ratio. Our focus is on the five, ten, and fifteen least important predictors, for which removal is a consideration.

We begin with investigating their aggregate contribution to the predictive performance of the *svm* and *logit* models. At each re-estimation date, we compute the fraction of aggregate absolute variable importance (i.e., the sum of decreases in likelihood ratio test statistic across all variables) that is attributed to these subsets of predictors. Illustratively for Germany, Panel B of Fig. 6 visualizes this fraction for the five (*bottom5*), ten (*bottom10*), and fifteen (*bottom15*) least important variables at each re-estimation date. Their average aggregate contribution remains relatively stable for all three subsets of predictor variables, i.e., on average, less important predictors remain less informative most of the time during the sample period. However, their contributions spike notably at multiple re-estimation dates, suggesting that even these seemingly unimportant covariate groups are material for both models (at least for some subperiods of the sample).

Further investigating the five least important predictors, we inspect the time variability in relative variable importance measures only within this subset of predictors. To this end, we omit the remaining variables prior to normalizing the absolute variable importance measures to sum to one at each re-estimation date. Panel C of Fig. 6 depicts the resulting relative variable importance metrics at each re-estimation date, using again Germany as an illustration. The lines indicate that the relative variable importance metrics fluctuate sharply over time. Therefore, we conclude that each predictor variable is an important contributor to the overall predictive power (albeit to varying degrees). The findings from Fig. 6 do *not* recommend we should remove specific predictors.[35]

6.2.3. Nonlinearity and interactions

Our results reveal that multivariate crash prediction models outperform their univariate counterparts because they incorporate information from multiple different predictors simultaneously to assess the overall economic conditions and to establish crash signals. In terms of classification performance, SVMs tend to outperform logistic regressions. They are constructed identically in terms of predictor variables, sample-splitting scheme, and re-estimation frequency, but they differ in their ability to capture nonlinearity and interactive effects. We thus inspect the potentially complex relationships between predictor variables and estimated crash probabilities, separately for the *svm* and *logit* models. Fig. 7 provides several instructive examples that document the importance of nonlinearity and interactions in establishing crash

[33] A *t*-test significantly rejects the null hypothesis that the correlation coefficients are zero (unreported).

[34] We simultaneously set the values for the one-month and one-year predictor pairs to their uninformative median values because those pairs are, by construction, highly correlated. We thus only show the results for the remaining twenty-two variables (omitting the six one-year predictors).

[35] To be on the conservative side, we compare the statistical and economic predictive performance of the original *svm* model with versions that only consider the top five, ten, or fifteen predictors in terms of their overall relative variable importance. Out-of-sample tests (unreported) are identical to the tests shown in Section 6.1. We find that no model version exhibits substantial outperformance in any of these tests, so we choose not to remove unconditionally less informative variables from the predictor set and instead consider each predictor as informative (albeit to varying degrees). Additionally, we caution that the pre-estimation variable selection based on relative importance metrics derived from the entire sample period could lead to foresight bias, undermining the credibility of any out-of-sample tests.

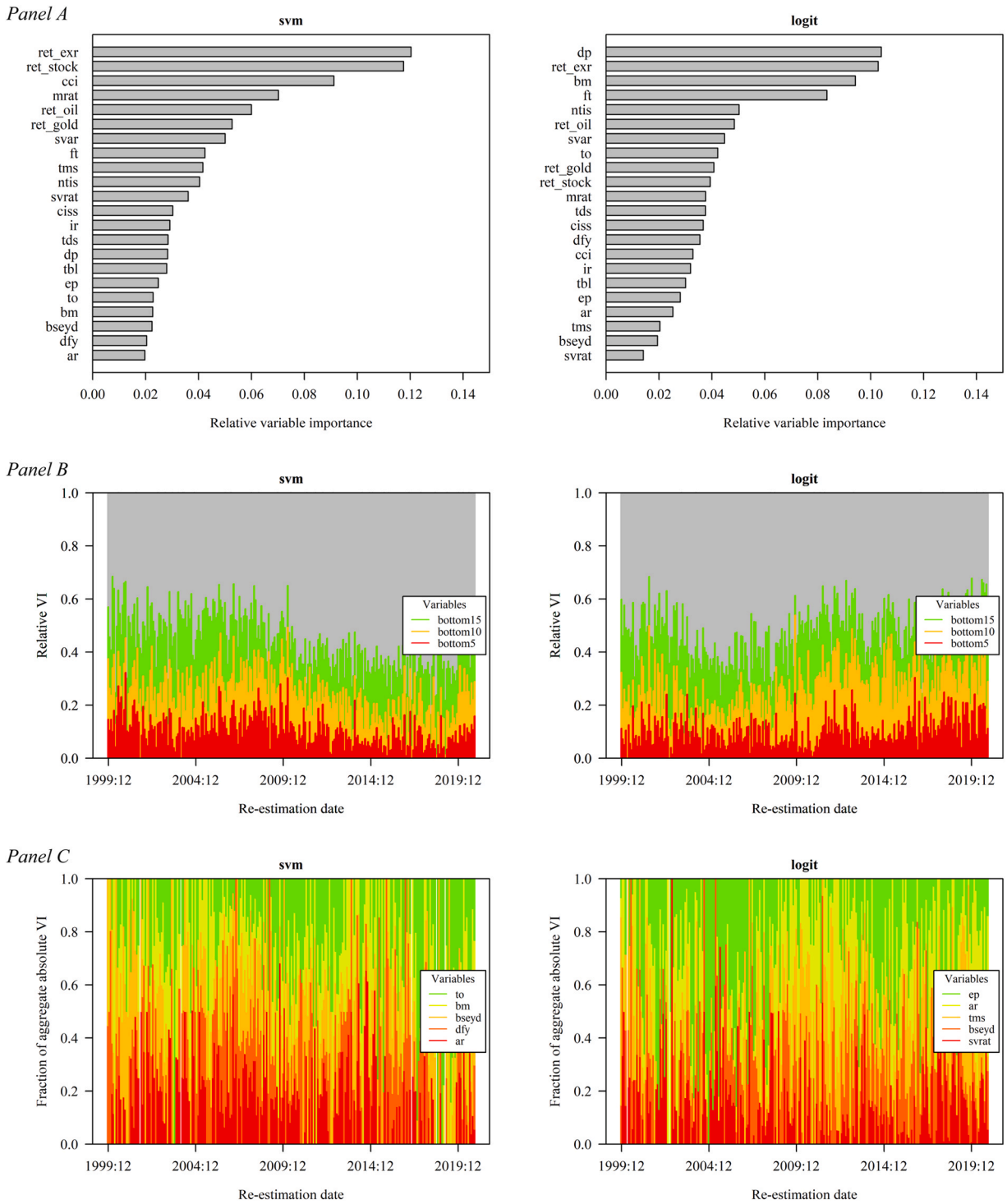

Fig. 6. Relative variable importance in aggregate and over time, Germany. This figure depicts relative variable importance metrics. Panel A presents the time series averages of relative variable importance measures during the January 2000–December 2020 out-of-sample period, which are calculated based on a two-step approach, separately for each re-estimation date: First, the absolute variable importance is computed as the decrease in likelihood ratio test statistic from setting all values of a given predictor to its uninformative median value within the balanced datasets (the values for the one-month and one-year predictor pairs are simultaneously set to their uninformative median values). Second, the absolute variable importance measures are normalized to sum to one, signaling the relative contribution of each variable to the respective model. Panel B visualizes the fraction of aggregate absolute variable importance (i.e., the sum of decreases in likelihood ratio test statistic across all variables) that is attributed to the five (*bottom5*), ten (*bottom10*), and fifteen (*bottom15*) least important variables at each re-estimation date. Focusing on the five least important predictors, Panel C presents the resulting relative variable importance metrics at each re-estimation date, but normalized within this subset of predictors. To this end, the remaining variables are omitted prior to normalizing the absolute variable importance measures to sum to one at each re-estimation date. The charts are presented for the multivariate crash prediction models based on logistic regressions and support vector machines (*logit* and *svm;* introduced in Section 5.2) illustratively for Germany. The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

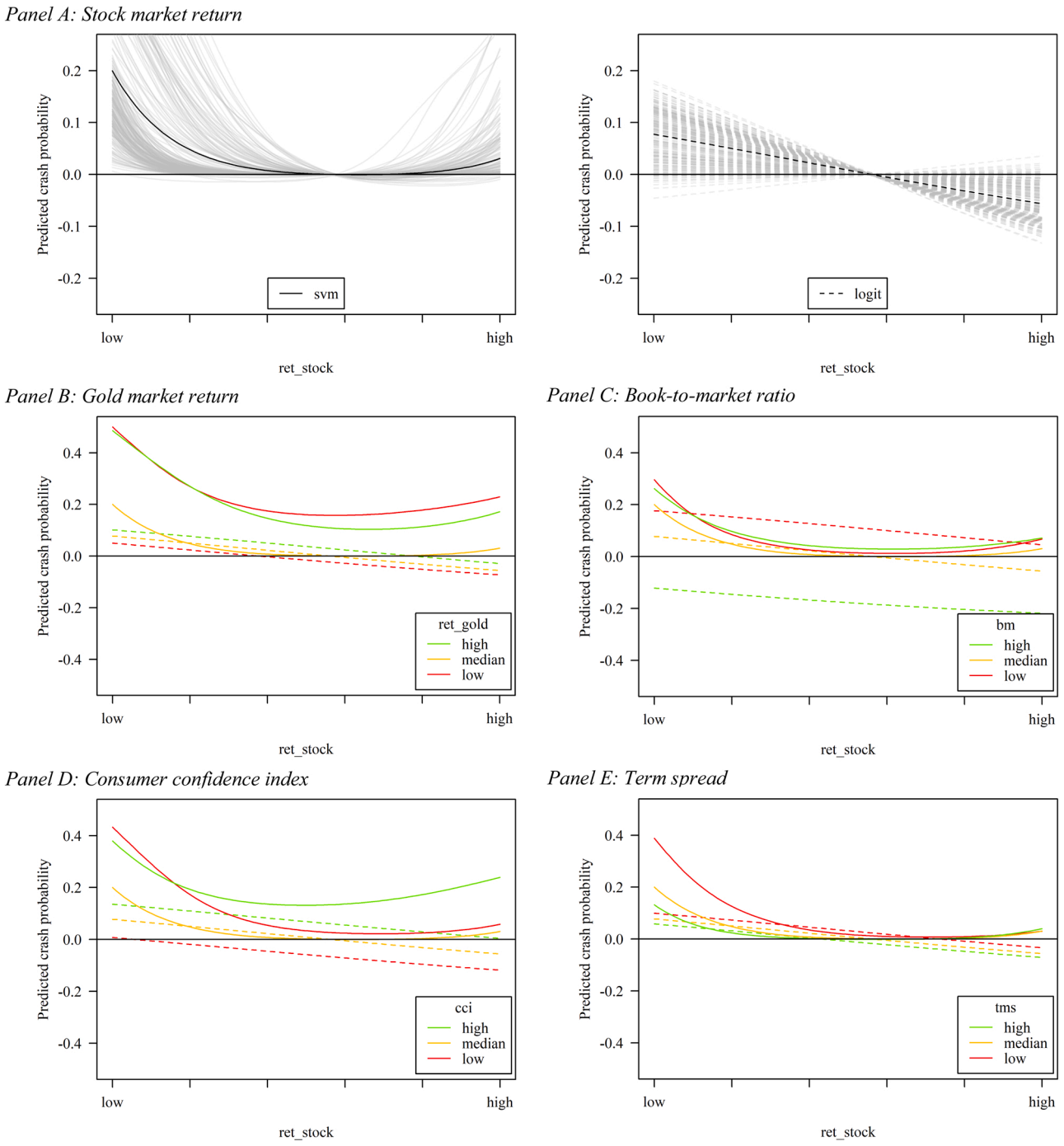

Fig. 7. Nonlinear and interactive effects in forecasting stock market crashes, Germany. This figure visualizes the nonlinear and interactive effects in the relationship between predictor variables and estimated crash probabilities. Panel A provides a first example. It presents the marginal effect of the current-month stock market return (*ret_stock*) on the predicted crash probability ($\overline{p}_{crash}$). To visualize the marginal effect of *ret_stock* on $\overline{p}_{crash}$, all predictors are set to their uninformative median values within the balanced datasets at each re-estimation date. *ret_stock* is then varied across the minimum and maximum values of its historical distribution and the change in $\overline{p}_{crash}$ relative to the median prediction is computed. The marginal association between *ret_stock* and $\overline{p}_{crash}$ is illustrated for each re-estimation date (grey lines) as well as averaged across all re-estimation dates (black lines). Panels B to E provide further examples by visualizing the interactive effects between *ret_stock* and the current-month gold market return (*ret_gold*, Panel B), the book-to-market ratio (*bm*, Panel C), the consumer confidence index (*cci*, Panel D), and the term spread (*tms*, Panel E) on $\overline{p}_{crash}$ using a similar approach. Replicating the procedure outlined above, the change in $\overline{p}_{crash}$ relative to the median prediction is computed for different levels of the four interaction variables, e.g., by varying *ret_gold* across its minimum and maximum values. The minimum and maximum levels for *ret_gold* and the other interaction variables are shown with a red and green line, respectively. The yellow line depicts the median values. The charts are presented for the multivariate crash prediction models based on logistic regressions and support vector machines (*logit* and *svm;* introduced in Section 5.2) illustratively for Germany. The figures use solid lines for support vector machines and dashed lines for logistic regressions. The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

signals. Again, we present and discuss the results illustratively for Germany.

In a first example, we examine the relationship between the current-month stock market return (*ret_stock*) and subsequent stock market crashes. To visualize the marginal effect of *ret_stock* on the predicted crash probability ($\overline{p}_{crash}$), we set all predictors to their uninformative median values within the balanced datasets at each re-estimation date. We then vary *ret_stock* across the minimum and maximum values of its historical distribution and compute the change in $\overline{p}_{crash}$ relative to the median prediction.[36] Positive (negative) values indicate an increase (decrease) in the predicted crash probability. Panel A of Fig. 7 illustrates the marginal association between *ret_stock* and $\overline{p}_{crash}$ for each re-estimation date (grey lines) as well as averaged across all re-estimation dates (black lines). The left figure contains the visualization for SVMs (solid lines), and the right figure that for logistic regressions (dashed lines). We find that the strength of the effects, i.e., the degree of skewness for the *svm* model or the magnitude of slope for the *logit* model, is weaker for some re-estimation dates and stronger for others. This supports our earlier observation that the underlying economic conditions that precede substantial stock market downturns are time-varying (see Fig. 2) and that each predictor variable's importance changes substantially over time (see Fig. 6, Panels B and C).

On average, we identify a negative quasi-linear relationship between *ret_stock* and $\overline{p}_{crash}$ for logistic regressions, and a non-symmetrical U-shaped relationship for SVMs.[37] During normal stock market regimes (i.e., when predictor values fluctuate around their historical median values), the predicted crash probability is close to the median prediction for both models. However, if stock market returns get abnormally low, the likelihood of a crash in the next month increases notably. This is in line with the literature suggesting that higher crash probabilities are caused by large changes in expected future cash flows or discount rates (Campbell and Shiller, 1988a, 1988b), stock market sentiment (Baker and Wurgler, 2006, 2007), or other changes in the underlying economic conditions that put pressure on stock markets beyond the current month. In addition, while the *logit* model indicates that larger-than-normal stock market returns decrease the estimated crash probability (relative to the median prediction), the *svm* model signals a slight increase in the estimated crash probability for abnormally high returns. This nonlinear effect is consistent with the literature on asset price bubbles, which build up after longer price run-ups and eventually burst (Brunnermeier, 2009; Greenwood et al., 2019). The example illustrates that SVMs identify a nonlinear relationship between *ret_stock* and $\overline{p}_{crash}$ that is consistent with economic theories, while logistic regressions fail to do so particularly at the right end of the crash probability distribution. It thus helps explain why the *svm* model outperforms the *logit* model (i.e., its linear multivariate benchmark).

Nevertheless, even extreme negative values for *ret_stock* increase the predicted crash probability relative to the median prediction by only around 20 percentage points (on average, across all re-estimation dates). Therefore, other predictor variables might simultaneously play important roles in increasing the predicted crash probability above the 50% threshold to signal a stock market crash during the next month. To investigate those determinants, and to illustrate in detail the benefits of SVMs relative to logistic regressions, we inspect the interactive effects between *ret_stock* and the current-month gold market return (*ret_gold*), the book-to-market ratio (*bm*), the consumer confidence index (*cci*), and the term spread (*tms*) on $\overline{p}_{crash}$. This analysis enables us to assess the underlying economic conditions from a price-based (Panel B), fundamentals-based (Panel C), sentiment-based (Panel D), and macroeconomic perspective (Panel E). We replicate the procedure outlined above. In this case, however, we compute the change in $\overline{p}_{crash}$ relative to the median prediction for different levels of the four interaction variables, e.g., by varying *ret_gold* across its minimum and maximum values. Panels B to E of Fig. 7 illustrate the interactive effects of *ret_stock* and *ret_gold/bm/cci/tms* on $\overline{p}_{crash}$. The minimum and maximum levels for *ret_gold* and the other interaction variables are shown with a red and green line, respectively. The yellow line depicts the median values. To keep the figures simple, we visualize the interactive effects averaged across all re-estimation dates. Again, we draw solid lines for SVMs and dashed lines for logistic regressions. A distance between the green/red line and the yellow line that varies over the crash probability distribution, e.g., a smaller or larger distance towards the end(s) of the distribution, indicates interactive effects.

As expected, in all Panels B to E, the yellow dashed lines for the *logit* model are only shifted up- or downward in a parallel way. In contrast, for the *svm* model, we uncover substantial interactive effects between *ret_stock* and *ret_gold/bm/cci/tms*, respectively. These effects are predominantly apparent for abnormally low stock market returns (at the left end of the distribution), whereas the right end of the distribution is less affected by the level of the interaction variables (as indicated by the smaller distance of the green and red lines from the yellow line).

The predicted crash probability notably increases (relative to the median prediction) when abnormally low stock market returns coincide with abnormally low or high gold market returns (*ret_gold*, Panel B). The first pattern refers to the "correlation breakdown" phenomenon. Correlations across asset classes increase substantially during strong market corrections (Longin and Solnik, 2001). Since reallocations of funds and sharp stock market corrections usually have a persistent market impact beyond the current month, and thus correlations remain elevated, the simultaneous occurrence of extreme negative stock and gold market returns may signal a stock market crash during the next month. The second pattern is consistent with a "safe haven" argument, suggesting that investors reallocate funds from stock markets to the gold market during crisis periods (Baur and McDermott, 2010). Gabaix and Koijen (2021) further show that the price elasticity of demand of the aggregate stock market is small, and even small outflows of funds from stock markets could cause large negative price reactions.

We identify similar interactive patterns for the book-to-market ratio (*bm*, Panel C) and the consumer confidence index (*cci*, Panel D). When abnormally low stock market returns coincide with market values that strongly differ from book values (i.e., abnormally low or high book-to-market ratios) or an abnormally low or high consumer confidence, these combinations lead to markedly higher predicted crash probabilities relative to the median predictions. A higher likelihood of a subsequent stock market crash in the case of low book-to-market ratios and high customer confidence is supported by behavioral theories stating that a bubble will eventually burst or deflate when it develops towards maturity and the investors' feedback trading loop is broken (De Long et al., 1990; Scherbina and Schlusche, 2014; Sornette and Cauwels, 2015). In contrast, higher crash probabilities in the case of high book-to-market ratios and low consumer confidence are supported by theories suggesting that ongoing stock market corrections are likely to persist beyond the current month, and that crashes are initiated and further fueled when the sentiment about a bubble asset is reversed (Barberis et al., 1998; Daniel et al., 1998; Scherbina and Schlusche, 2014).

Finally, when stock market returns are abnormally low in an environment of flattening or even reversion yield curves, i.e., when long-term interest rates get close to or even fall below the short-term

[36] Instead of pre-determining a standard deviation-based departure from the mean or a multiple of the interquartile range, our approach captures outliers that are necessary for extreme event predictions, while maintaining a realistic range of values.

[37] Logistic regressions incorporate linear relationships between predictors and the log-odds that a crash will occur during the next month. This leads to an S-shaped relationship between predictor variables and estimated crash probabilities (bounded at 0 and 1), which shows up as a quasi-linear relationship around the center of the distribution (between these two boundaries). We also identify *ret_gold/bm/cci/tms* and $\overline{p}_{crash}$ for the *logit* model, and non-symmetrical U-shaped relationships for the *svm* model (unreported).

interest rates, the predicted crash probability sizably increases (relative to the median prediction). This pattern is consistent with evidence showing that abnormally low or even negative term spreads possess significant predictive power for economic recessions (Estrella and Hardouvelis, 1991; Rudebusch and Williams, 2009) and declines in consumption growth (Harvey, 1988), which in turn adversely affect stock markets (Ferson and Harvey, 1993).

These instructive examples help explain why the *svm* model outperforms the *logit* model (i.e., its linear multivariate benchmark). SVMs incorporate relevant interactions inherently. Interacting *ret_stock* with other predictor variables increases the predicted crash probability relative to the median prediction significantly more than in the case when considering *ret_stock* in isolation. Logistic regressions are unable to capture these effects if no *pre-determined* terms are added to the regression model, although the functional form of those effects is generally unknown ex ante. We conclude that abnormally low stock market returns in the current month do not necessarily point towards an even stronger stock market correction in the next month, *unless* other economic indicators are simultaneously falling out of their normal ranges (e.g., the yield curve flattens or even reverses).

In summary, nonlinear and interactive effects provide an explanation for the advantageousness of multivariate machine learning-based stock market crash prediction models over their univariate counterparts and, more importantly, for the advantageousness of SVMs over logistic regressions. The use of machine learning techniques further enables us to uncover robust statistical relationships in the underlying economic conditions that precede stock market crashes. Our results thus confirm Lopez de Prado's (2020) conclusion that machine learning should not prematurely be regarded as a black box.

7. Conclusion

Using a comprehensive set of twenty-eight price-based, fundamentals-based, sentiment-based, and macroeconomic predictor variables from the five largest Eurozone countries, we compare the performance of simple univariate and machine learning-based multivariate models in forecasting subsequent stock market crashes. We show that there is no single variable or small subset of variables that always and reliably precedes stock market crashes with extreme values, suggesting substantial time variation in the predictive ability of any single variable. Multivariate crash prediction models should be advantageous over their univariate counterparts because they are capable of incorporating the information content of multiple predictor variables simultaneously. Our results support this view. In terms of statistical predictive performance, a support vector machine-based crash prediction model outperforms a random classifier and is superior to the average univariate benchmark. It also performs better than a multivariate logistic regression model, which is unable to capture nonlinearity and interactive effects. We provide several instructive examples to demonstrate that incorporating nonlinear and interactive effects is both imperative and foundation for the outperformance of support vector machines.

Our findings contribute to the early-warning literature and have two important implications: First, an accurate model for predicting stock market crashes out-of-sample translates into substantial value-added to active investors. Second, for policymakers, the use of machine learning-based crash prediction models can help activate macroprudential policy tools in time, maybe in combination with other models from the early-warning literature, aiming at increasing resiliency of the financial system as a whole and mitigating the imminent costs of financial crises.

Appendix A

Following Lleo and Ziemba (2017), we evaluate a crash prediction model's predictive performance using the conditional probability $P(CI = 1|CS = 1)$, i.e., the probability that stock market crashes actually occur when they are expected to occur. We first compare the two components, the sequence of crash signals $CS := \{CS_{t+1|t} \in \{1,0\}; t = 1,\ldots,T\} = \{CS_1,\ldots,CS_T\}$ and the sequence of crash indicators $CI := \{CI_{t+1} \in \{1,0\}; t = 1,\ldots,T\} = \{CI_1,\ldots,CI_T\}$. In each month $t+1$, the binary variables equal 1 if a crash is expected to occur $(CS_{t+1|t} = 1)$ or actually occurs $(CI_{t+1} = 1)$, and 0 otherwise. We then create a monthly hit indicator X_i, which equals 1 for true positive classifications, and 0 for false positive classifications:

$$X_i = \begin{cases} 1 \text{ if } CI_{t+1} = 1 | CS_{t+1|t} = 1 \\ 0 \text{ if } CI_{t+1} = 0 | CS_{t+1|t} = 1 \end{cases}. \tag{6}$$

The length of the hit indicator sequence $X := \{X_i \in \{1,0\} | i = 1,\ldots,N\} = \{X_1,\ldots,X_N\}$ shrinks to the number of distinct crash signals $N = \sum_{t=1}^{T} CS_{t+1|t}$. The conditional random variable X follows a Bernoulli distribution with probability $p = P(CI = 1|CS = 1)$, which can be estimated by using the maximum likelihood estimator $\widehat{p} = \frac{\sum_{i=1}^{N} X_i}{N}$. It maximizes the log-likelihood function $l(p|X) = \ln L(p|X) = \sum_{i=1}^{N} X_i \ln(p) + (N - \sum_{i=1}^{N} X_i)\ln(1-p)$ based on the likelihood function $L(p|X) = \prod_{i=1}^{N} p^{X_i}(1-p)^{1-X_i}$, and represents a crash prediction model's overall precision, namely the number of true positive classifications divided by the number of distinct crash signals (Lleo and Ziemba, 2017).

We also follow Lleo and Ziemba (2017) in testing whether a crash prediction model's classification performance is significantly different from that of a random classifier. The no-information rate, i.e., the conditional probability that random signals correctly forecast crashes, reflects the number of crashes that occurred during the sample period divided by the number of sample months: $p_0 = \frac{\sum_{t=1}^{T} CI_{t+1}}{T}$.[38] Testing the null hypothesis that a crash prediction model's conditional probability is equal to the no-information rate, the likelihood ratio test statistic is:

$$Y = -2\ln(\Lambda), \text{ with } \Lambda = \frac{L(p = p_0|X)}{L(p = \widehat{p}|X)}, \tag{7}$$

for which we report the empirical p-statistics, corrected for a potential small sample bias stemming from the low number of crash signals.[39]

[38] For example, sixteen crashes occurred in Germany during the sample period (252 months). Therefore, the no-information rate, which equals the historical crash probability, is $p_0 = \frac{16}{252} \approx 6.35\%$.

[39] In general, the Y-statistic is asymptotically χ^2-distributed with $\nu = 1$ degree of freedom. But as the χ^2 distribution is continuous and only valid asymptotically, it may not provide an adequate approximation for the discrete empirical distribution of test statistics stemming from the low number of crash signals. To correct for this potential small sample bias, we follow Ziemba, Zhitlukhin, and Lleo (2017) and compute the empirical p-statistics from 10,000 Monte Carlo simulations.

Appendix B

Table B1
Statistical predictive performance (additional sample countries).

Model specification		Classification metrics												
Model	Sign	n_{preds}	TP	FP	TN	FN	Acc. [%]	Prec. [%]	Rec. [%]	F1 [%]	*Y*-stat	*p*-stat [%]	CP1	CP2	CP3
France ($n_{periods} = 252$, $n_{crashes} = 11$)															
Panel A: Univariate crash models ($K_{exit} = 95\%$, $K_{entry} = 95\%$)															
bseyd	–	18	4	14	227	7	92	22	36	28	7.23	0.07	✓	×	×
ep	+	19	4	15	226	7	91	21	36	27	6.83	0.11	✓	×	×
dp	–	21	4	17	224	7	90	19	36	25	6.12	0.20	✓	×	×
svar	+	13	3	10	231	8	93	23	27	25	5.64	0.22	✓	✓	×
ft	+	9	2	7	234	9	94	22	18	20	3.62	0.56	×	✓	✓
svrat	+	13	2	11	230	9	92	15	18	17	2.35	1.69	×	✓	✓
ret_ri_ann	–	15	2	13	228	9	91	13	18	15	1.91	2.55	✓	×	×
tds	+	17	2	15	226	9	90	12	18	14	1.55	3.48	✓	✓	×
ret_exr_ann	+	17	2	15	226	9	90	12	18	14	1.55	3.48	✓	×	×
mrat	–	19	2	17	224	9	90	11	18	13	1.26	47.33	✓	×	×
ret_gold_ann	+	25	2	23	218	9	87	8	18	11	0.64	41.24	✓	×	×
tms	–	10	1	9	232	10	92	10	9	10	0.56	70.72	✓	×	×
ret_exr	–	10	1	9	232	10	92	10	9	10	0.56	70.72	×	✓	×
ret_oil_ann	+	13	1	12	229	10	91	8	9	8	0.28	66.29	×	✓	×
ret_ri	–	14	1	13	228	10	91	7	9	8	0.22	65.69	×	×	✓
ciss	–	33	2	31	210	9	84	6	18	9	0.20	39.31	✓	×	×
cci	–	16	1	15	226	10	90	6	9	7	0.12	63.36	✓	×	×
bm	+	17	1	16	225	10	90	6	9	7	0.08	63.06	✓	×	×
dfy	+	18	1	17	224	10	89	6	9	7	0.06	62.63	×	✓	×
ar	+	18	1	17	224	10	89	6	9	7	0.06	62.63	×	×	✓
svar_ann	+	28	1	27	214	10	85	4	9	5	0.04	62.62	✓	×	×
ntis	–	19	1	18	223	10	89	5	9	7	0.03	62.44	✓	×	×
to	–	27	1	26	215	10	86	4	9	5	0.03	62.16	✓	×	×
ir	+	12	0	12	229	11	91	0	0	0			×	×	×
ir_ann	+	12	0	12	229	11	91	0	0	0			×	×	×
ret_gold	+	16	0	16	225	11	89	0	0	0			×	×	×
ret_oil	+	15	0	15	226	11	90	0	0	0			×	×	×
tbl	+	11	0	11	230	11	91	0	0	0			×	×	×
Panel B: Multivariate crash models															
logit		14	3	11	230	8	92	21	27	24	5.22	0.28	✓	✓	✓
svm		**21**	**4**	**17**	**224**	**7**	**90**	**19**	**36**	**25**	**6.12**	**0.20**	✓	✓	✓

Model specification		Classification metrics												
Model	Sign	n_{preds}	TP	FP	TN	FN	Acc. [%]	Prec. [%]	Rec. [%]	F1 [%]	*Y*-stat	*p*-stat [%]	CP1	CP2	CP3
Italy ($n_{periods} = 252$, $n_{crashes} = 11$)															
Panel A: Univariate crash models ($K_{exit} = 95\%$, $K_{entry} = 95\%$)															
ret_exr	–	10	3	7	234	8	94	30	27	29	7.20	0.05	×	✓	×
to	–	21	4	17	224	7	90	19	36	25	6.12	0.20	✓	✓	×
bseyd	–	14	3	11	230	8	92	21	27	24	5.22	0.28	×	✓	×
ep	+	17	3	14	227	8	91	18	27	21	4.19	0.55	×	✓	×
ciss	+	18	3	15	226	8	91	17	27	21	3.91	0.62	×	✓	×
ft	+	9	2	7	234	9	94	22	18	20	3.62	0.56	×	✓	✓
bm	+	21	3	18	223	8	90	14	27	19	3.17	1.14	✓	✓	×
svrat	+	12	2	10	231	9	92	17	18	17	2.61	1.34	×	✓	✓
ret_ri_ann	–	13	2	11	230	9	92	15	18	17	2.35	1.69	✓	×	×
tds	+	17	2	15	226	9	90	12	18	14	1.55	3.48	✓	✓	×
ret_oil_ann	–	17	2	15	226	9	90	12	18	14	1.55	3.48	✓	✓	×
dfy	+	18	2	16	225	9	90	11	18	14	1.40	48.69	×	✓	×
mrat	–	18	2	16	225	9	90	11	18	14	1.40	48.69	✓	×	×
dp	–	34	3	31	210	8	85	9	27	13	1.26	27.46	✓	×	×
svar_ann	+	20	2	18	223	9	89	10	18	13	1.13	46.12	×	✓	×
ret_ri	–	10	1	9	232	10	92	10	9	10	0.56	70.72	×	✓	×
tms	–	10	1	9	232	10	92	10	9	10	0.56	70.72	✓	×	×
ntis	+	12	1	11	230	10	92	8	9	9	0.36	67.48	✓	×	×
svar	+	13	1	12	229	10	91	8	9	8	0.28	66.29	×	✓	×
ret_gold	–	14	1	13	228	10	91	7	9	8	0.22	65.69	×	✓	×
ret_exr_ann	–	14	1	13	228	10	91	7	9	8	0.22	65.69	×	✓	×
cci	+	17	1	16	225	10	90	6	9	7	0.08	63.06	×	✓	×
ar	+	18	1	17	224	10	89	6	9	7	0.06	62.63	×	×	✓
ret_gold_ann	+	25	1	24	217	10	87	4	9	6	0.01	61.78	✓	×	×
ir	+	8	0	8	233	11	92	0	0	0			×	×	×
ir_ann	+	8	0	8	233	11	92	0	0	0			×	×	×
ret_oil	+	15	0	15	226	11	90	0	0	0			×	×	×
tbl	+	11	0	11	230	11	91	0	0	0			×	×	×

(*continued on next page*)

Table B1 (*continued*)

Model specification		Classification metrics												
Model	Sign	n_{preds}	TP	FP	TN	FN	Acc. [%]	Prec. [%]	Rec. [%]	F1 [%]	*Y*-stat	*p*-stat [%]	CP1	CP2	CP3
Panel B: Multivariate crash models															
logit		27	4	23	218	7	88	15	36	21	4.45	0.56	×	✓	✓
svm		**32**	**4**	**28**	**213**	**7**	**86**	**13**	**36**	**19**	**3.44**	**1.14**	✓	✓	✓
Model specification		Classification metrics													
Model	Sign	n_{preds}	TP	FP	TN	FN	Acc. [%]	Prec. [%]	Rec. [%]	F1 [%]	*Y*-stat	*p*-stat [%]	CP1	CP2	CP3
Spain ($n_{periods} = 252$, $n_{crashes} = 11$)															
Panel A: Univariate crash models ($K_{exit} = 95\%$, $K_{entry} = 95\%$)															
bm	+	21	4	17	224	7	90	19	36	25	6.12	0.20	✓	✓	×
to	–	22	4	18	223	7	90	18	36	24	5.80	0.25	✓	✓	×
svrat	–	15	3	12	229	8	92	20	27	23	4.85	0.35	×	✓	×
ep	+	19	3	16	225	8	90	16	27	20	3.64	0.81	✓	✓	×
ft	+	9	2	7	234	9	94	22	18	20	3.62	0.56	×	✓	✓
ret_exr	–	10	2	8	233	9	93	20	18	19	3.23	0.79	×	✓	×
cci	–	21	3	18	223	8	90	14	27	19	3.17	1.14	✓	✓	×
tds	+	17	2	15	226	9	90	12	18	14	1.55	3.48	✓	✓	×
ret_oil_ann	–	17	2	15	226	9	90	12	18	14	1.55	3.48	✓	✓	×
ciss	+	18	2	16	225	9	90	11	18	14	1.40	48.69	×	✓	×
dfy	+	18	2	16	225	9	90	11	18	14	1.40	48.69	×	✓	×
svar_ann	+	19	2	17	224	9	90	11	18	13	1.26	47.33	×	✓	×
bseyd	–	21	2	19	222	9	89	10	18	13	1.01	44.88	×	✓	×
ntis	–	24	2	22	219	9	88	8	18	11	0.72	41.97	✓	×	×
dp	–	24	2	22	219	9	88	8	18	11	0.72	41.97	✓	×	✓
svar	+	10	1	9	232	10	92	10	9	10	0.56	70.72	×	✓	×
tms	–	10	1	9	232	10	92	10	9	10	0.56	70.72	✓	×	×
tbl	+	11	1	10	231	10	92	9	9	9	0.45	69.04	✓	×	×
ret_ri_ann	–	11	1	10	231	10	92	9	9	9	0.45	69.04	✓	×	×
ret_exr_ann	–	14	1	13	228	10	91	7	9	8	0.22	65.69	×	✓	×
mrat	–	16	1	15	226	10	90	6	9	7	0.12	63.36	×	✓	×
ar	+	18	1	17	224	10	89	6	9	7	0.06	62.63	×	×	✓
ret_gold_ann	+	25	1	24	217	10	87	4	9	6	0.01	61.78	✓	×	×
ir	+	5	0	5	236	11	94	0	0	0			×	×	×
ir_ann	+	4	0	4	237	11	94	0	0	0			×	×	×
ret_gold	+	16	0	16	225	11	89	0	0	0			×	×	×
ret_oil	+	15	0	15	226	11	90	0	0	0			×	×	×
ret_ri	+	11	0	11	230	11	91	0	0	0			×	×	×
Panel B: Multivariate crash models															
logit		30	1	29	212	10	85	3	9	5	0.08	63.34	×	✓	×
svm		**30**	**3**	**27**	**214**	**8**	**86**	**10**	**27**	**15**	**1.69**	**29.44**	×	✓	✓
Model specification		Classification metrics													
Model	Sign	n_{preds}	TP	FP	TN	FN	Acc. [%]	Prec. [%]	Rec. [%]	F1 [%]	*Y*-stat	*p*-stat [%]	CP1	CP2	CP3
Netherlands ($n_{periods} = 252$, $n_{crashes} = 15$)															
Panel A: Univariate crash models ($K_{exit} = 95\%$, $K_{entry} = 95\%$)															
svar	+	11	3	8	229	12	92	27	20	23	5.02	0.32	✓	✓	✓
mrat	–	19	4	15	222	11	90	21	27	24	4.86	0.46	✓	×	×
ret_ri	–	14	3	11	226	12	91	21	20	21	3.73	0.77	✓	✓	✓
cci	–	14	3	11	226	12	91	21	20	21	3.73	0.77	✓	×	✓
ft	+	9	2	7	230	13	92	22	13	17	2.61	1.34	×	✓	✓
dp	–	28	4	24	213	11	86	14	27	19	2.55	20.05	✓	×	✓
ret_ri_ann	–	18	3	15	222	12	89	17	20	18	2.55	2.05	✓	✓	×
tms	–	10	2	8	229	13	92	20	13	16	2.26	1.88	✓	×	×
ret_exr	–	10	2	8	229	13	92	20	13	16	2.26	1.88	×	✓	×
svrat	+	13	2	11	226	13	90	15	13	14	1.47	48.53	×	✓	✓
ret_oil_ann	+	13	2	11	226	13	90	15	13	14	1.47	48.53	✓	✓	×
to	–	13	2	11	226	13	90	15	13	14	1.47	48.53	✓	×	×
tds	+	17	2	15	222	13	89	12	13	13	0.81	42.17	✓	✓	×
ret_exr_ann	+	17	2	15	222	13	89	12	13	13	0.81	42.17	✓	×	×
ciss	+	18	2	16	221	13	88	11	13	12	0.69	41.14	×	✓	×
svar_ann	–	18	2	16	221	13	88	11	13	12	0.69	41.14	✓	×	×
bm	+	21	2	19	218	13	87	10	13	11	0.41	39.24	✓	×	×
bseyd	–	21	2	19	218	13	87	10	13	11	0.41	39.24	✓	✓	×
ep	+	23	2	21	216	13	87	9	13	11	0.27	38.88	✓	✓	×
ret_gold_ann	+	25	2	23	214	13	86	8	13	10	0.17	73.65	✓	×	×
ir	+	14	1	13	224	14	89	7	7	7	0.03	62.10	✓	×	×
ir_ann	+	14	1	13	224	14	89	7	7	7	0.03	62.10	✓	×	×
ret_gold	–	14	1	13	224	14	89	7	7	7	0.03	62.10	×	✓	×
ret_oil	–	15	1	14	223	14	89	7	7	7	0.01	61.80	×	×	✓
dfy	+	18	1	17	220	14	88	6	7	6	0.01	61.42	×	✓	×
ar	+	18	1	17	220	14	88	6	7	6	0.01	61.42	×	×	✓
ntis	–	18	1	17	220	14	88	6	7	6	0.01	61.42	×	✓	×
tbl	+	11	0	11	226	15	90	0	0	0			×	×	×

(*continued on next page*)

Table B1 (*continued*)

Model specification		Classification metrics												
Model	Sign	n_{preds}	TP	FP	TN	FN	Acc. [%]	Prec. [%]	Rec. [%]	F1 [%]	*Y*-stat	*p*-stat [%]	CP1	CP2	CP3
Panel B: Multivariate crash models															
logit		17	5	12	225	10	91	29	33	31	9.09	0.02	✓	✓	✓
svm		**31**	**7**	**24**	**213**	**8**	**87**	**23**	**47**	**30**	**9.33**	**0.03**	✓	✓	✓

This table reports the classification metrics for different crash prediction models for France, Italy, Spain, and the Netherlands during the January 2000–December 2020 out-of-sample period. The classification metrics are based on the comparison of $CS_{t+1|t}$ and CI_t. CI_{t+1} is a binary *crash indicator*, which equals 1 when a substantial stock market downturn occurs during month $t+1$, and 0 otherwise. $CS_{t+1|t}$ is a binary *crash signal*, which equals 1 if the respective model, incorporating all information available at the end of month *t*, expects a crash to occur during month $t+1$, and 0 otherwise. Panel A presents the metrics for the univariate crash prediction models introduced in Section 5.1. In addition to the number of crash signals (n_{preds}), the numbers of true/false positives (*#TP/#FP*) and true/false negatives (*#TN/#FN*) are reported, together with accuracy (Acc.), precision (Prec.), recall (Rec.), and F1 measures. The likelihood ratio test statistic (*Y*-stat) testing the null hypothesis that a crash prediction model's conditional probability is equal to the no-information rate is added, together with the empirical *p*-statistic (*p*-stat), as well as the distribution of true positives across the three major crash periods (CP1/CP2/CP3). The last three columns indicate whether the respective model is able to correctly forecast at least one stock market crash within each of the three subperiods of the sample (surrounding the three major crash periods), i.e., January 2000-December 2007 (CP1), January 2008-December 2014 (CP2), and January 2015-December 2020 (CP3). Panel B presents the metrics for the multivariate crash prediction models based on logistic regressions and support vector machines (*logit* and *svm*; introduced in Section 5.2). The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

Table B2
Statistical predictive performance (additional crash prediction models), Germany.

Model specification		Classification metrics													
Model	Sign	n_{preds}	TP	FP	TN	FN	Acc. [%]	Prec. [%]	Rec. [%]	F1 [%]	*Y*-stat	*p*-stat [%]	CP1	CP2	CP3
Germany ($n_{periods} = 252$, $n_{crashes} = 16$)															
Panel A: Univariate crash models ($K_{exit} = 95\%$, $K_{entry} = 90\%$)															
ep	+	21	5	16	220	11	89	24	31	27	6.61	0.16	✓	×	×
dfy	+	23	5	18	218	11	88	22	31	26	5.84	0.29	✓	✓	×
mrat	–	23	5	18	218	11	88	22	31	26	5.84	0.29	✓	✓	×
ret_exr	–	10	3	7	229	13	92	30	19	23	5.24	0.28	×	✓	×
ft	+	11	3	8	228	13	92	27	19	22	4.70	0.41	×	✓	✓
ret_ri_ann	–	20	4	16	220	12	89	20	25	22	4.14	0.67	✓	✓	×
bm	+	30	5	25	211	11	86	17	31	22	3.81	15.09	✓	×	×
svar	+	15	3	12	224	13	90	20	19	19	3.10	1.21	✓	✓	×
ret_ri	–	15	3	12	224	13	90	20	19	19	3.10	1.21	✓	✓	✓
dp	–	34	5	29	207	11	84	15	31	20	2.98	12.70	✓	×	✓
ret_exr_ann	+	26	4	22	214	12	87	15	25	19	2.62	20.09	✓	×	×
ret_gold_ann	+	38	5	33	203	11	83	13	31	19	2.31	11.29	✓	×	✓
tds	–	30	4	26	210	12	85	13	25	17	1.91	17.96	✓	×	×
svar_ann	+	21	3	18	218	13	88	14	19	16	1.68	28.68	×	✓	✓
bseyd	–	21	3	18	218	13	88	14	19	16	1.68	28.68	✓	×	×
svrat	+	12	2	10	226	14	90	17	13	14	1.53	48.90	×	✓	✓
ret_oil_ann	–	22	3	19	217	13	87	14	19	16	1.51	27.77	✓	✓	×
ciss	+	24	3	21	215	13	87	13	19	15	1.21	26.66	×	✓	×
cci	–	14	2	12	224	14	90	14	13	13	1.12	45.04	✓	✓	×
ir	–	27	1	26	210	15	84	4	6	5	0.37	41.21	×	✓	×
ir_ann	–	27	1	26	210	15	84	4	6	5	0.37	41.21	×	✓	×
tms	–	11	1	10	226	15	90	9	6	7	0.12	63.04	✓	×	×
ar	+	26	2	24	212	14	85	8	13	10	0.07	72.58	×	×	✓
to	–	12	1	11	225	15	90	8	6	7	0.07	62.48	✓	×	×
ntis	–	36	2	34	202	14	81	6	13	8	0.04	72.04	✓	×	✓
tbl	+	17	1	16	220	15	88	6	6	6	0.01	61.34	✓	×	×
ret_gold	–	15	1	14	222	15	88	7	6	6	0.00	61.15	×	✓	×
ret_oil	–	15	1	14	222	15	88	7	6	6	0.00	61.15	×	✓	×
Panel B: Multivariate crash models															
logit		34	5	29	207	11	84	15	31	20	2.98	12.70	✓	✓	✓
svm		**29**	**7**	**22**	**214**	**9**	**88**	**24**	**44**	**31**	**9.43**	**0.03**	✓	✓	✓
rf		27	6	21	215	10	88	22	38	28	7.23	0.13	✓	✓	✓
gbrt		28	6	22	214	10	87	21	38	27	6.87	0.16	✓	✓	✓
nn		36	8	28	208	8	86	22	50	31	9.64	0.02	✓	✓	✓

This table reports the classification metrics for different crash prediction models illustratively for Germany during the January 2000–December 2020 out-of-sample period. The classification metrics are based on the comparison of $CS_{t+1|t}$ and CI_t. CI_{t+1} is a binary *crash indicator*, which equals 1 when a substantial stock market downturn occurs during month $t+1$, and 0 otherwise. $CS_{t+1|t}$ is a binary *crash signal*, which equals 1 if the respective model, incorporating all information available at the end of month *t*, expects a crash to occur during month $t+1$, and 0 otherwise. Panel A presents the metrics for the univariate crash prediction models introduced in Section 5, considering an alternative assumption for the exit and entry thresholds: $K_{exit} = 95\%$ and $K_{entry} = 90\%$. In addition to the number of crash signals (n_{preds}), the numbers of true/false positives (*#TP/#FP*) and true/false negatives (*#TN/#FN*) are reported, together with accuracy (Acc.), precision (Prec.), recall (Rec.), and F1 measures. The likelihood ratio test statistic (*Y*-stat) testing the null hypothesis that a crash prediction model's conditional probability is equal to the no-information rate is added, together with the empirical *p*-statistic (*p*-stat), as well as the distribution of true positives across the three major crash periods (CP1/CP2/CP3). The last three columns indicate whether the respective model is able to correctly forecast at least one stock market crash within each of the three subperiods of the sample (surrounding the three major crash periods), i.e., January 2000-December 2007 (CP1), January 2008-December 2014 (CP2), and January 2015-December 2020 (CP3). Panel B presents the metrics for the multivariate crash prediction models based on logistic regressions and support vector machines (*logit* and *svm;* introduced in Section 5.2) as well as random

forests (*rf*), gradient boosted regression trees (*gbrt*), and neural networks (*nn*). The sample includes the five largest Eurozone countries by gross domestic product as of December 2019 (Germany, France, Italy, Spain, and the Netherlands) during the January 1990–December 2020 sample period. The data coming from Refinitiv are collected on a monthly basis and, if currency-related, denominated in Euro.

Appendix C

Table C1
Stylized visualization | Binary classifications.

Panel A					Crash indicator	
					1 (crash occured)	0 (no crash occured)
			Crash signal	1 (crash predicted)	True positives (TPs)	False positive (FPs)
				0 (no crash predicted)	False negatives (FNs)	True negative (TNs)

Panel B	Confusion matrix				Classification measures			
	Realized							
		1	0		Accuracy	Precision	Recall	F1
Predicted	1	20	10	Definition	$\frac{\#(TP+TN)}{\#(TP+FP+FN+TN)}$	$\frac{\#TP}{\#(TP+FP)}$	$\frac{\#TP}{\#(TP+FN)}$	$2\times\frac{Precision\times Recall}{Precision+Recall}$
	0	5	15	Calculation	$\frac{20+15}{20+10+5+15}=70\%$	$\frac{20}{20+10}=67\%$	$\frac{20}{20+5}=80\%$	$2\times\frac{67\%\times 80\%}{67\%+80\%}=73\%$

This table depicts two stylized visualizations that help explain the procedure to measure the performance of binary classifications.

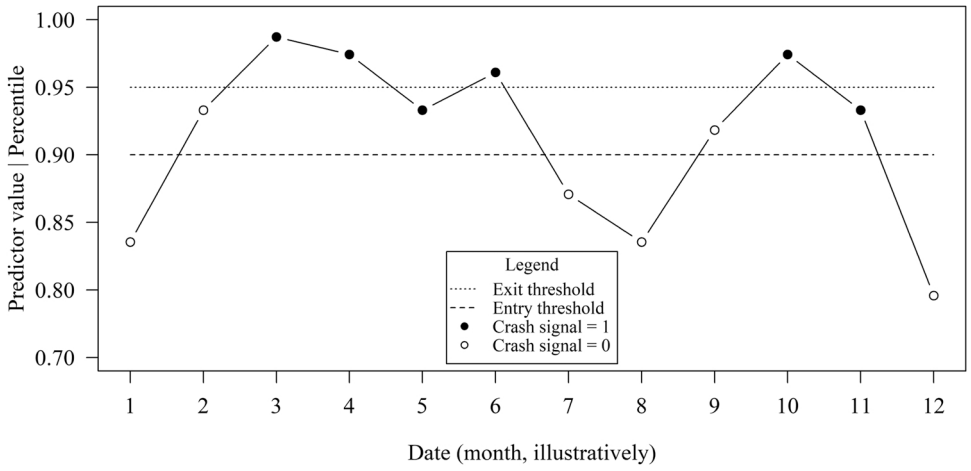

Fig. C1. Stylized visualization | Univariate crash prediction models. This figure depicts a stylized visualization that helps explain the structure and functioning of univariate crash prediction models.

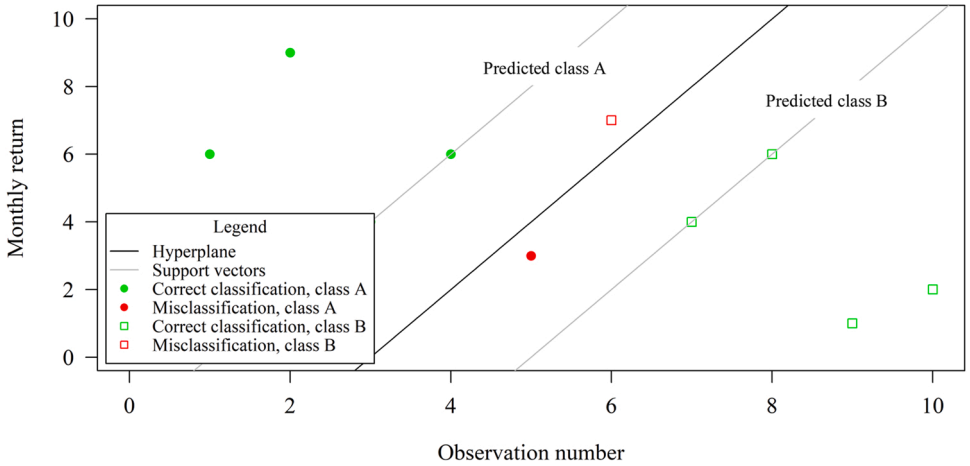

Fig. C2. Stylized visualization | Support vector machines. This figure depicts a stylized visualization that helps explain the structure and functioning of support vector machines (SVMs) in a two-dimensional, two-class scenario. Each vector (observation) is defined by two variables and assigned to one of the two classes, and the SVM searches for a hyperplane (a straight line) that territorially divides the vector space (the two-dimensional shape) into groups of vectors that belong to the same class by aiming to 1) maximize the distance of correctly classified support vectors from the hyperplane and 2) minimize the number of misclassified support vectors .

References

Alessi, L., Detken, C., 2011. Quasi real time early warning indicators for costly asset price boom/bust cycles: a role for global liquidity. Eur. J. Political Econ. 27 (3), 520–533.

Alessi, L., Detken, C., 2018. Identifying excessive credit growth and leverage. J. Financ. Stab. 35 (1), 215–225.

Aliber, R.Z., Kindleberger, C.P., 2015. Manias, Panics, and Crashes: A History of Financial Crises. Palgrave Macmillan, Basingstoke, U.K.

Ang, A., Bekaert, G., 2004. How regimes affect asset allocation. Financ. Anal. J. 60 (2), 86–99.

Ang, A., Timmermann, A., 2012. Regime changes and financial markets. Annu. Rev. Financ. Econ. 4 (1), 313–337.

Ang, A., Goyal, A., Ilmanen, A., 2014. Asset allocation and bad habits. Rotman Int. J. Pension Manag. 7 (2), 16–27.

Angel, J.J., Broms, T.J., Gastineau, G.L., 2016. ETF transaction costs are often higher than investors realize. J. Portf. Manag. 42 (3), 65–75.

Asness, C., 2003. Fight the fed model. J. Portf. Manag. 30 (1), 11–24.

Avramov, D., Kaplanski, G., Subrahmanyam, A., 2020. Moving average distance as a predictor of equity returns. Rev. Financ. Econ. 39 (2), 127–145.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. J. Financ. 61 (4), 1645–1680.

Baker, M., Wurgler, J., 2007. Investor sentiment in the stock market. J. Econ. Perspect. 21 (2), 129–152.

Bandopadhyaya, A., Jones, A.L., 2008. Measures of investor sentiment: a comparative analysis put-call ratio vs. volatility index. J. Bus. Econ. Res. 6 (8), 27–34.

Barberis, N., Shleifer, A., Vishny, R., 1998. A model of investor sentiment. J. Financ. Econ. 49 (3), 307–343.

Baron, M., Verner, E., Xiong, W., 2021. Banking crises without panics. Q. J. Econ. 136 (1), 51–113.

Barro, R.J., Ursúa, J.F., 2017. Stock-market crashes and depressions. Res. Econ. 71 (3), 384–498.

Baur, D.G., McDermott, T.K., 2010. Is gold a safe haven? International evidence. J. Bank. Financ. 34 (8), 1886–1898.

Berge, K., Consigli, G., Ziemba, W.T., 2008. The predictive ability of the bond-stock earnings yield differential model. J. Portf. Manag. 34 (3), 63–80.

Beutel, J., List, S., von Schweinitz, G., 2019. Does machine learning help us predict banking crises? J. Financ. Stab. 45 (1), 100693.

Billingsley, R.S., Chance, D.M., 1988. Put-call ratios and market timing effectiveness. J. Portf. Manag. 15 (1), 25–28.

Bluwstein, K., Buckmann, M., Joseph, A., Kang, M., Kapadia, S., and Şimşek, Ö. (2020). Credit Growth, the Yield Curve and Financial Crisis Prediction: Evidence from a Machine Learning Approach. BoE Staff Working Paper No. 848.

Borkovec, M., Serbin, V., 2013. Create or buy: a comparative analysis of liquidity and transaction costs for selected U.S. ETFs. J. Portf. Manag. 39 (4), 118–131.

Brunnermeier, M.K., Oehmke, M., 2013. Bubbles, financial crises, and systemic risk. In: Constantinides, G.M., Harris, M., Stulz, R.M. (Eds.), Handbook of the Economics of Finance. Elsevier, Amsterdam, Netherlands.

Brunnermeier, M.K., Vernengo, M., Caldentey, E.P., Rosser Jr., B.J., 2009. Bubbles. The New Palgrave Dictionary of Economics. Palgrave, London, U.K.

Bulla, J., Mergner, S., Bulla, I., Sesboüé, A., Chesneau, C., 2011. Markov-switching asset allocation: do profitable strategies exist? J. Asset Manag. 12 (5), 310–321.

Campbell, J.Y., Shiller, R.J., 1988a. The dividend-price ratio and expectations of future dividends and discount factors. Rev. Financ. Stud. 1 (3), 195–228.

Campbell, J.Y., Shiller, R.J., 1988b. Stock prices, earnings, and expected dividends. J. Financ. 43 (3), 661–676.

Cenesizoglu, T., Timmermann, A., 2012. Do return prediction models add economic value? J. Bank. Financ. 36 (11), 2974–2987.

Chatzis, S.P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., Vlachogiannakis, N., 2018. Forecasting stock market crisis events using deep and statistical machine learning techniques. Expert Syst. Appl. 112 (1), 353–371.

Choudhry, R., Garg, K., 2008. A hybrid machine learning system for stock market forecasting. Int. J. Comput. Inf. Eng. 2 (3), 689–692.

Chow, G., Jacquier, E., Kritzman, M., Lowry, K., 1999. Optimal portfolios in good times and bad. Financ. Anal. J. 55 (3), 65–73.

Claessens S., and Kose, A. (2013). Financial Crisis: Explanations Types, and Implications. IMF Working Paper No. 2013/028.

Copeland, M.M., Copeland, T.E., 1999. Market timing: style and size rotation using the VIX. Financ. Anal. J. 55 (2), 73–81.

Daniel, K., Hirshleifer, D., Subrahmanyam, A., 1998. Investor psychology and security market under- and overreactions. J. Financ. 53 (6), 1839–1885.

De Long, J.B., Shleifer, A., Summers, L., Waldmann, R., 1990. Noise trader risk in financial markets. J. Political Econ. 98 (4), 703–738.

Dichtl, H., Drobetz, W., Kryzanowski, L., 2016. Timing the stock market: does it really make no sense? J. Behav. Exp. Financ. 10 (1), 88–104.

Dietterich, T., 2000. Ensemble methods in machine learning. Lecture Notes in Computer Science: Multiple Classifier Systems. Springer, Berlin, Germany.

Drobetz, W., Otto, T., 2021. Empirical asset pricing via machine learning: evidence from the european stock market. J. Asset Manag. 22 (7), 507–538.

Estrada, J., 2006. The fed model: a note. Financ. Res. Lett. 3 (1), 14–22.

Estrella, A., Hardouvelis, G., 1991. The term structure as a predictor of real economic activity. J. Financ. 46 (2), 555–576.

Fama, E., 2014. Two pillars of asset pricing. Am. Econ. Rev. 104 (6), 1467–1485.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (1), 3133–3181.

Ferrer, E., Salaber, J., Zalewska, A., 2016. Consumer confidence indices and stock markets' meltdowns. Eur. J. Financ. 22 (3), 195–220.

Ferson, W., Harvey, C.R., 1993. The risk and predictability of international equity returns. Rev. Financ. Stud. 6 (3), 527–566.

Fischer, S., Merton, R.C., 1984. Macroeconomics and finance: the role of the stock market. Carne -Rochester Conf. Ser. Public Policy 21 (1), 57–108.

Fouliard, J., Howell, M., and Rey, H. (2021). Answering the Queen: Machine Learning and Financial Crises. NBER Working Paper No. 28302.

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Rev. Financ. Stud. 33 (5), 2326–2377.

Gabaix, X., and Koijen, R.S. (2021). In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis. Swiss Finance Institute Research Paper No. 20–91.

Goetzmann, W.N., Kim, D., 2018. Negative bubbles: what happens after a crash. Eur. Financ. Manag. 24 (2), 171–191.

Goetzmann, W.N., Kim, D., and Shiller, R.J. (2017). Crash Beliefs from Investor Surveys. NBER Working Paper No. 22143.

Gorton, G., 2018. Financial crises. Annu. Rev. Financ. Econ. 10 (1), 43–58.

Goyal, G., Ilmanen, A., Kabiller, D., 2015. Bad habits and good practices. J. Portf. Manag. 41 (4), 97–107.

Greenwood, R., Shleifer, A., You, Y., 2019. Bubbles for Fama. J. Financ. Econ. 131 (1), 20–43.

Greenwood, R., Hanson, S., Shleifer, A., Sorensen, J., 2021. Predictable financial crises. J. Financ. 77 (2), 863–921.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financ. Stud. 33 (5), 2223–2273.

Harvey, C.R., 1988. The real term structure and consumption growth. J. Financ. Econ. 22 (2), 305–334.

Holló, D., Kremer, M., and Lo Duca, M. (2012). CISS – A Composite Indicator of Systemic Stress in The Financial System. ECB Working Paper No. 1426.

Huang, L.H., Chang, Y.C., 2022. Growth impact of equity market crisis: a global perspective. Int. Rev. Econ. Financ. 78 (1), 154–176.

Huang, W., Nakamori, Y., Wang, S.-Y., 2005. Forecasting stock market movement direction with support vector machines. Comput. Oper. Res. 32 (10), 2513–2522.

Jones, B. (2016). Institutionalizing Countercyclical Investment: A Framework for Long-term Asset Owners. IMF Working Paper No. 2016/038.

Jordà, Ò., Schularick, M., Taylor, A.M., 2017. Macrofinancial history and the new business cycle facts. NBER Macroecon. Annu. 31 (1), 213–263.

Kaminsky, G., Reinhart, C., 1999. The twin crisis: the causes of banking and balance-of-payments problems. Am. Econ. Rev. 89 (3), 473–500.

Kelly, B.T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: a unified model of risk and return. J. Financ. Econ. 134 (3), 501–524.

Khatibi, V., Khatibi, E., Rasouli, A., 2011. A new support vector machine-genetic algorithm (SVM-GA) based method for stock market forecasting. Int. J. Phys. Sci. 6 (25), 6091–6097.

Kritzman, M., Li, Y., 2010. Skulls, financial turbulence, and risk management. Financ. Anal. J. 44 (5), 30–41.

Kritzman, M., Li, Y., Page, S., Rigobon, R., 2011. Principal components as a measure of systemic risk. J. Portf. Manag. 37 (4), 112–126.

Laeven, L., Valencia, F., 2020. Systemic banking crises database II. IMF Econ. Rev. 68 (2), 307–361.

Lee, T.K., Cho, J.H., Kwon, D.S., Sohn, S.Y., 2019. Global stock market investment strategies based on financial network indicators using machine learning techniques. Expert Syst. Appl. 117 (1), 228–242.

Leitch, G., Tanner, J.E., 1991. Economic forecast evaluation: profits versus the conventional error measures. Am. Econ. Rev. 81 (3), 580–590.

Leung, M., Daouk, H., Chen, A.-S., 2000. Forecasting stock indices: a comparison of classification and level estimation models. Int. J. Forecast. 16 (2), 173–190.

Lewellen, J., 2015. The cross-section of expected stock returns. Crit. Financ. Rev. 4 (1), 1–44.

Lleo, S., Ziemba, W.T., 2017. Does the bond-stock earnings yield differential model predict equity market corrections better than high P/E models? Financ. Mark. 26 (2), 61–123.

Lleo, S., Ziemba, W.T., 2019. Can Warren Buffett forecast equity market corrections? Eur. J. Financ. 25 (4), 369–393.

Longin, F., Solnik, B., 2001. Extreme correlation of international equity markets. J. Financ. 56 (2), 649–676.

Lopez de Prado, M.M. (2020). Machine Learning for Asset Managers. Cambridge, U.K.: Cambridge Elements (Quantitative Finance).

Lucas, R. (1976). Macroeconomic Policy Evaluation: A Critique. Carnegie-Rochester Conference Series on Public Policy, 1(2), 19–46.

Luque, A., Carrasco, A., Martín, A., de las Heras, A., 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit. 91 (1), 216–231.

Mahalanobis, P.C., 1936. On the generalised distance in statistics. Proc. Natl. Inst. Sci. India 49–55.

Maio, P., 2013. The fed model and the predictability of stock returns. Rev. Financ. 17 (4), 1489–1533.

McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? J. Financ. 71 (1), 5–32.

McPhillips, L.E., Chang, H., Chester, M.V., Depietri, Y., Friedman, E., Grimm, N.B., Kominoski, J.S., McPhearson, T., Méndez-Lázaro, P., Rosi, E.J., Shiva, J.S., 2018. Defining extreme events: a cross-disciplinary review. Earth'S. Future 6 (3), 441–455.

Menardi, G., Torelli, N., 2014. Training and assessing classification rules with imbalanced data. Data Min. Knowl. Discov. 28 (1), 92–122.

Neely, C.J., Rapach, D.E., Tu, J., Zhou, G., 2014. Forecasting the equity risk premium: the role of technical indicators. Manag. Sci. 60 (7), 1617–1859.

Ni, L.-P., Ni, Z.-W., Gao, Y.-Z., 2011. Stock trend prediction based on fractal feature selection and support vector machine. Expert Syst. Appl. 38 (5), 5569–5576.
Nystrup, P., Hansen, B.W., Madsen, H., Lindström, E., 2015. Regime-based versus static asset allocation: letting the data speak. J. Portf. Manag. 42 (1), 103–109.
Ohana, J.J., Ohana, S., Benhamou, E., Saltiel, D., and Guez, B. (2021). Explainable AI Models of Stock Crashes: A Machine-Learning Explanation of the Covid March 2020 Equity Meltdown. Working Paper.
Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. J. Account. Res. 18 (1), 109–131.
Pastor, L., Veronesi, P., 2006. Was there a Nasdaq bubble in the late 1990s? J. Financ. Econ. 81 (1), 61–100.
Pastor, L., Veronesi, P., 2009. Technological revolutions and stock prices. Am. Econ. Rev. 99 (4), 1713–1757.
Reinhart, C.M., Reinhart, V.R., 2015. Financial crisis, development and growth: a long-term perspective. World Bank Econ. Rev. 29 (1), 53–76.
Reinhart, C.M., Rogoff, K., 2011. From financial crash to debt crisis. Am. Econ. Rev. 101 (5), 1676–1706.
Ren, R., Wu, D.D., Liu, T., 2019. Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Syst. J. 13 (1), 760–770.
Rudebusch, G., Williams, J., 2009. Forecasting recessions: the puzzle of the enduring power of the yield curve. J. Bus. Econ. Stat. 27 (4), 492–503.
Samitas, A., Kampouris, E., Kenourgios, D., 2020. Machine learning as an early warning system to predict financial crisis. Int. Rev. Financ. Anal. 71 (1), 101507.
Scherbina, A., Schlusche, B., 2014. Asset price bubble: a survey. Quant. Financ. 14 (4), 589–604.
Schularick, M., Taylor, A.M., 2012. Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870-2008. Am. Econ. Rev. 102 (2), 1029–1061.
Shao, Y., Lunetta, R.S., 2012. Comparison of support vector machine, neural network, and cart algorithms for the land-cover classification using limited training data points. ISPRS J. Photogramm. Remote Sens. 70 (1), 78–87.
Shiller, R.J., 2003. From efficient markets theory to behavioral finance. J. Econ. Perspect. 17 (1), 83–104.
Shiryaev, A.N., Zhitlukhin, M.V., Ziemba, W.T., 2014. When to sell apple and the Nasdaq? Trading bubbles with a stochastic disorder model. J. Portf. Manag. 40 (2), 54–63.
Sornette, D., Cauwels, P., 2015. Financial bubbles: mechanisms and diagnostics. Rev. Behav. Econ. 2 (3), 279–305.
Tanaka, K., Kinkyo, T., Hamori, S., 2016. Random forests-based early warning system for bank failures. Econ. Lett. 148 (1), 118–121.
Vapnik, V.N., 1998. The support vector method of function estimation. J. A. Suykens, and J. Vandewalle, Nonlinear Modeling. Springer, Boston (MA), U.S.
Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21 (4), 1455–1508.
West, K.D., 2006. Handbook of Economic Forecasting. In: Elliott, G., Granger, C., Timmermann, A. (Eds.). Elsevier, Amsterdam, Netherlands.
Whaley, R.E., 2000. The investor fear gauge. J. Portf. Manag. 26 (3), 12–17.
Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Trans. Evolut. Comput. 1 (1), 67–82.
Yu, L., Wang, S., Lai, K.K., 2005. A novel adaptive learning algorithm for stock market prediction. Lect. Notes Comput. Sci. 3827 (1), 443–452.
Yu, L., Chen, H., Wang, S., Lai, K.K., 2009. Evolving least squares support vector machines for stock market trend mining. IEEE Trans. Evolut. Comput. 13 (1), 87–102.
Ziemba, W.T., and Schwartz, S. (1991). Invest Japan: The Structure, Performance and Opportunities of Japan's Stock, Bond and Fund Markets. Chicago (IL), U.S.: Probus.
Ziemba, W.T., Zhitlukhin, M., Lleo, S., 2017. Stock Market Crashes: Predictable and Unpredictable and What to do About Them. Republic of Singapore: World Scientific Publishing, Singapore.