International Journal of Finance & Economics

WILEY

SPECIAL ISSUE ARTICLE

Enhancing Market Predictability and Investment Decision-Making: Machine Learning Models for Predicting Stock Market Crashes in China

Mingtao Zhou[1] | Yong Ma[2] | He Ni[3]

[1]School of Economics and Management, Changsha University of Science and Technology, Changsha, China | [2]College of Finance and Statistics, Hunan University, Changsha, China | [3]Tailong Finance School, Zhejiang Gongshang University, Hangzhou, China

Correspondence: He Ni (nihe@zjgsu.edu.cn)

Received: 19 June 2025 | **Revised:** 25 September 2025 | **Accepted:** 18 October 2025

Funding: This work was supported by National Natural Science Foundation of China (Grant 72371098) and Zhejiang Provincial Philosophy and Social Sciences Planning Project (Grant 26NDJC082YB).

Keywords: crash determinants | investment decision | machine learning | model evaluation | stock market crashes

ABSTRACT

This study aims to enhance the predictive accuracy of China's stock market crashes and optimise the investment decision-making process by leveraging machine learning techniques and a diverse array of constructed aggregate factors. The empirical analysis validates the superior market-timing capabilities of machine learning tools, particularly the feedforward neural network, in comparison to the conventional logistic regression model. This advancement, in turn, translates into nontrivial economic benefits for mean–variance investors across various portfolio types, even after accounting for trading costs. Moreover, our findings illuminate that stocks with small capitalisations, traded on the Shenzhen Stock Exchange, and within aggressive industry sectors exhibit heightened sensitivity to market crashes. Additionally, we identify substantial supplementary effects of aggregate market-level characteristics in conjunction with well-established macro factors. Notably, momentum, valuation and liquidity emerge as the most influential market-level feature groups in predicting market crashes.

JEL Classification: C52, C53, G01, G11

1 | Introduction

The examination of stock price behaviour and the pursuit of reliable predictions in stock markets has been an enduring area of interest. Extensive research has reported the potential predictability of equity risk premium (see Campbell and Thompson 2008; Rapach et al. 2013, for recent surveys). In contrast, limited studies are devoted to the periods of time when the equity market encounters rare black swan events (i.e., crash periods) (Nyberg 2013). Despite the rarity of these crash events, they often lead to substantial economic, social, and political costs. For instance, the 2015 Chinese stock market erased nearly 30% of the Shanghai Composite Index within a few weeks, posing severe threats to financial system stability and investor confidence.

Early detection of crashes is therefore crucial. For policymakers, reliable forecasts allow timely contingency planning to mitigate the imminent costs associated with market crashes (Thorbecke 1997; Bjørnland and Leitemo 2009). For investors, they enable informed risk management and asset allocation (Tu 2010). Accordingly, our study aims to develop an effective out-of-sample prediction model for Chinese stock market crashes by systematically evaluating a broad set of machine learning techniques and identifying the best-performing approach. The recent surge in computational capabilities has

fuelled a wave of applications of machine learning in finance and economics, yielding encouraging results (Khandani et al. 2010; Gu et al. 2020; Dichtl et al. 2023). Machine learning methods are particularly appealing because of their efficacy in handling high-dimensional data and nonlinear dynamics that traditional econometric tools struggle to address. These strengths are especially relevant in the Chinese context, where frequent structural breaks, episodic regulatory interventions, and the dominance of retail investors introduce complex market dynamics and a rich set of predictive signals. Therefore, in such an environment, the more flexible machine learning methods are required to account for the Chinese market's specificity.

Our evaluation considers two benchmark generalised linear models: a logistic regression model using well-known macroeconomic variables (LOGIT-M); a logistic regression model with all predictors (LOGIT), and six machine learning methods: principal component analysis (PCA); regularised logistic regression with the elastic net penalty term (ENet); linear discriminant analysis (LDA); gradient boosted classification trees (GBCT); support vector machine (SVM); feed-forward neural network (NN).[1]

Through various tests, we first report the out-of-sample performance results and demonstrate that machine learning models, especially the neural network, unambiguously improve the predictability of stock market crashes when compared with traditional LOGIT models. In contrast with linear machine learning techniques, we confirm that nonlinear supervised learning methodologies, like GBCT and NN, further improve the prediction of stock market crashes, implying the complexity of China's stock market and emphasising the necessity of adopting highly flexible methods to account for intricate nonlinearities and interactions in the data. Moreover, the statistical test results show that the neural network model produces a maximum number of significant pairs and emerges as the best-performing model in terms of statistical significance.

Second, we delve into the practical implications of the generated predictions, assessing their impact on the improvement of investment decisions and their economic relevance in guiding portfolio allocations. Our findings reveal the notable efficacy of machine learning methods in supporting the investment decision-making process. Specifically, the long-short portfolios comprising all A-share stocks, guided by machine learning predictions, achieve cumulative returns ranging between 2.42 and 5.67, all of which are higher than those produced by the long-short portfolios with conventional logistic regression models' predictions. Notably, the neural network model outperforms its competitors, yielding the highest cumulative return for both the long-short strategy (5.67) and the long-bond strategy (3.27). These results underscore the potential economic benefits of leveraging advanced machine learning methods, particularly the neural network model, within the realm of financial decision-making.

Third, our results unveil the heterogeneous portfolio performances across different market segments, industry sectors, and individual stocks. In particular, our findings indicate that stocks traded on the Shenzhen Stock Exchange, within highly aggressive industries and with small capitalisations, outperform their corresponding counterparts. This discernible divergence in performance implies the varying sensitivity among different types of stocks to stock market crashes, contributing to a nuanced understanding of the differential impacts of market crashes on diverse segments of the stock market. Meanwhile, all machine learning models, especially the neural network, prove robust to these different portfolio types in terms of investment gains. This robustness further underscores the adaptability and efficacy of the neural network model in capturing China's stock market dynamics.

Finally, we strive to enhance our understanding of which type of characteristics is more critical in predicting stock market crashes. It is demonstrated that market-level information exhibits strong supplementary effects to widely used macroeconomic factors in predicting China's stock market crashes. This is not accidental. The macroeconomic characteristics contain relatively small market-specific information about the future trajectory of the stock market due to the heterogeneity of participants' expectations. Furthermore, we find that almost all machine learning methods agree on a similar set of principal market-level variable groups. The first group is related to the market price trend. The second group consists of valuation ratios. Lastly, liquidity risk signals are also among the dominant variable groups.

Our study extends the literature in three primary ways. First, we broaden applications of machine learning in finance by systematically evaluating machine learning methods for early detection of China's stock market crashes—an area where prior work has relied primarily on generalised linear models (e.g., Bussiere and Fratzscher 2006; Chevapatrakul 2013; Zhang et al. 2019; Fu et al. 2020). Second, we link statistical improvements to economic value by assessing the portfolio implications of machine learning-based forecasts, addressing the gap between predictive accuracy and economic benefits (Leitch and Tanner 1991). Third, we enrich the China-specific crash literature by constructing a comprehensive set of monthly market-level signals, complementing the long-standing focus on macroeconomic predictors related to output, production, employment, money and rates (e.g., Estrella and Mishkin 1998; Harvey 1988; Chen 2009) whose market-wide forecasting power is limited (Roll 1988; Morck et al. 2000). Besides, we apply the advanced SHapley Additive exPlanations (SHAP) AI technique to uncover the relative importance and influential mechanisms of determinants in predicting Chinese stock market crashes, thereby providing a more nuanced understanding of crash drivers and offering guidance for policymakers seeking targeted interventions to mitigate market crashes.

The remainder of the article is organised as follows. Section 2 describes the methodologies and evaluation metrics used for prediction and assessment. Section 3 outlines the data and variables. Section 4 reports the model performance. Section 5 delves further into the investment decision implications of machine learning predictions. Section 6 examines the forecasting power of a variety of predictors. Section 7 concludes.

2 | Methodology and Evaluation Metrics

2.1 | Methodology

Throughout our analysis, the relation between the stock market regimes (i.e., crash and non-crash) and their corresponding predictors is defined as:

$$S_{t+1} = \mathbb{E}_t\left[S_{t+1}\right] + \epsilon_{t+1} \tag{1}$$

where

$$\mathbb{E}_t\left[S_{t+1}\right] = F\left(x_t\right) \tag{2}$$

is the month-t expected market state; S_{t+1} is a binary variable that takes the value of 1 (a crash period) or 0 (a non-crash period); $F(\cdot)$ is a functional of a vector of characteristics x_t.

In this study, we follow a conventional procedure for hyperparameter optimisation, model training, and performance evaluation. Specifically, we split the data into disjoint training (Jan 2002–Dec 2011) and testing (Jan 2012–Jun 2021) periods.[2] The training set is used to tune hyperparameters and select the decision threshold, while the testing set always consists of the month immediately following the current training window.[3] Models are reestimated each month in an expanding-window scheme to produce one-step-ahead forecasts. For example, we first train on data from January 2002 through December 2011 to predict the regime in January 2012; then we expand the training window to January 2002 through January 2012 to predict the regime in February 2012. Because crashes are rare, training on the raw data would bias estimation toward the majority (non-crash) class. We therefore mitigate class imbalance by (i) using stratified 3-fold cross-validation with time-contiguous folds that preserve the crash/non-crash ratio within the training window and (ii) applying random oversampling of the minority (crash) class only within the training folds to avoid data leakage.[4]

Moreover, similar to Gu et al. (2020), to ensure the stability and robustness of the results, we adopt an ensemble method to train our machine learning models for two primary reasons. First, it controls the stochastic nature of oversampling, cross-validation, and optimisation in the model training process. Second, it reduces the dispersion and variance of model predictions and overcomes the local optimum trap. In particular, we use multiple random states to implement our machine learning estimation and derive predictions by selecting the mode of forecasts from a machine learning algorithm with 10 initial values.

2.2 | Evaluation Metrics

Because crash prediction is a rare-event, imbalanced classification problem, naive accuracy can be misleading. For example, suppose there are 100 months, of which only 5 are crash months and 95 are non-crash months. A trivial classifier that always predicts 'non-crash' attains 95% accuracy, yet it misses every crash and is therefore inappropriate for our study. Accordingly, we report four complementary metrics that are widely used for the evaluation of imbalanced classification. Based on the 2×2 confusion matrix $\mathcal{M} = \begin{bmatrix} \mathrm{TP} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TN} \end{bmatrix}$, where TP (TN) is the number of correctly identified crash (non-crash) months and FP (FN) is the number of months incorrectly labelled as crash (non-crash), the metrics are:

$$F_1\,\mathrm{score} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

$$G\text{-mean} = \sqrt{\frac{\mathrm{TP} \cdot \mathrm{TN}}{(\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP})}} \tag{4}$$

$$N\varphi = \frac{1}{2}\left(\frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}} + 1\right) \tag{5}$$

$$N\kappa = \frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP})}{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{FP} + \mathrm{TN}) + (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{FN} + \mathrm{TN})} + \frac{1}{2} \tag{6}$$

The F_1-score balances precision $\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\right)$ and recall $\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\right)$, aligning with the goal of detecting crashes (high recall) without excessive false alarms (high precision). G-mean balances sensitivity $\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\right)$ and specificity $\left(\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right)$, discouraging models that perform well on the abundant non-crash class but poorly on the scarce crash class. The Matthews correlation coefficient ($N\varphi$) uses all four cells of the confusion matrix and remains informative under class imbalance, providing a single summary of overall discriminatory power. Cohen's kappa ($N\kappa$) adjusts for chance agreement, guarding against spuriously high scores from trivially labelling most months as non-crash. Together, these measures provide a comprehensive assessment tailored to rare-event detection in financial markets.

3 | Data and Variable Definitions

This section starts by defining the stock market crash in China and then describes the key predictors used in our empirical analysis. Specifically, our primary source of stock market data is the Wind Economic Database. The macroeconomic variables are collected from their authoritative data publishers.[5] The sample period spans from January 2002 to June 2021.[6]

3.1 | Definition of Stock Market Crash

It seems reasonable to characterise a stock market crash as a steep or persistent drop in the overall equity market index (Coudert and Gex 2008). To this end, we introduce a quantified indicator, $CMAX_t$, which is first designed by Patel and Sarkar (1998) and further developed in studies of either detecting extreme price shocks or dating equity market risks over a given period of time (see e.g., Coudert and Gex 2008; Leroy and Pop 2019; Lu et al. 2021). Specifically, this involves dividing the month-t stock market index P_t by the maximum index value over the past fixed-length period $\mathcal{L}$:

$$CMAX_t = \frac{P_t}{\max\left(P_{t-\mathcal{L}}, \cdots, P_t\right)} \tag{7}$$

where $\mathcal{L}$ is set to 1 year (12 months) following Fu et al. (2020), Lu et al. (2021) and Li et al. (2022). This should be sensible since the lengthy window will give rise to unexpected information loss. $CMAX_t$ equals one if the equity market index rises over the past 12 months. The more the equity market index declines, the closer $CMAX_t$ gets to zero. Subsequently, we recognise crash

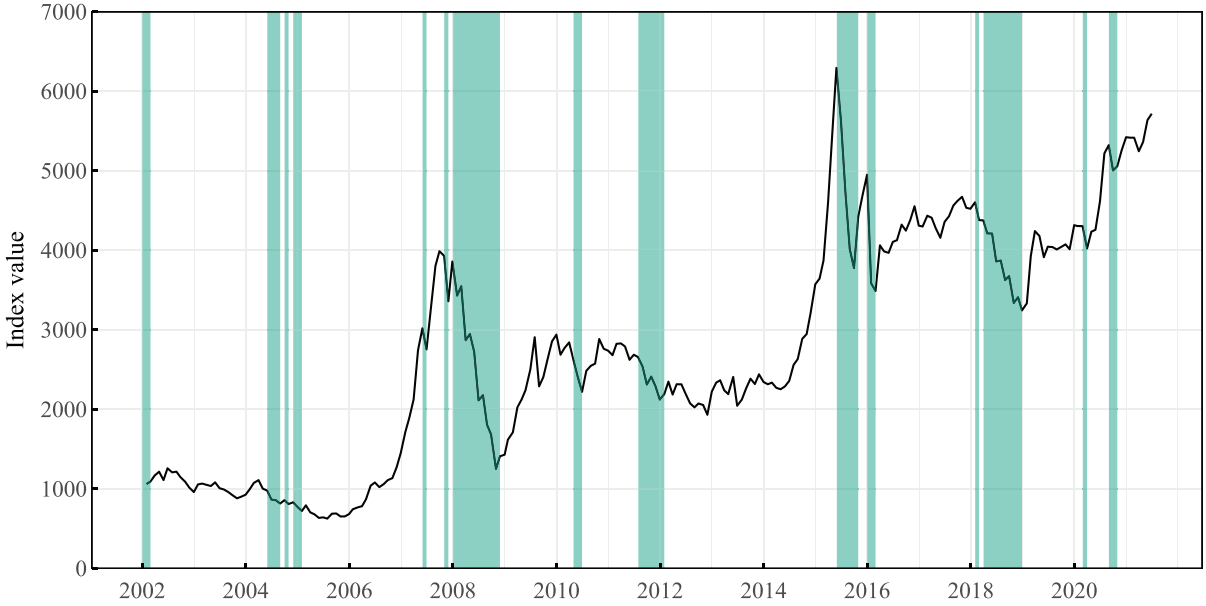

FIGURE 1 | Decoding states of China's stock market. This figure displays the evolution of the transition between the different market states for China's stock market with reference to the Wind All A-share index. The green background represents the crash periods and the white background stands for the non-crash periods. The sample period is from January 2002 to June 2021. [Colour figure can be viewed at wileyonlinelibrary.com]

periods with abnormally low $CMAX_t$ ratio by applying a threshold which is usually equal to one or two standard deviations below the mean.

By setting a threshold with a mean less than one standard deviation, we identify crises consistent with recognised crash events over the sample period in China's stock market.[7] Specifically, the stock market crash indicator I_t is defined as:

$$I_t = \begin{cases} 1, & CMAX_t < \mu_t - \sigma_t \\ 0, & CMAX_t \geq \mu_t - \sigma_t \end{cases} \tag{8}$$

where μ_t and σ_t are the mean and the standard deviation of $CMAX_t$, respectively. They are rolling calculated over the past 12 months.

In our study, we identify the stock market crash with reference to the Wind All A-share Index for avoiding the sample selection bias. It is a free-floating weighted stock market index of all A-share stocks listed on both the Shanghai Stock Exchange and the Shenzhen Stock Exchange in China, making it reasonable to comprehensively reflect the overall stock market fluctuations in China. Compared with the Wind All A-share Index, other commonly used indexes are based on a part of the stocks traded in China's stock market. For example, the Shanghai Stock Exchange Composite Index only accounts for stocks listed on the Shanghai Stock Exchange but ignores stocks traded at the Shenzhen Stock Exchange. In fact, as the two main stock exchanges in China, their stock distribution and trading volume are considerably divergent. Also, the widely used CSI 300 Index is designed to replicate the performance of the top 300 constituent stocks in China's stock market exchanges.

Figure 1 displays the dynamic evolution of the transition between different market states for China's stock market with respect to the Wind All A-share index over the sample period (January 2002–June 2021). The state decoding result is well consistent with our understanding of China's stock market fluctuations and successfully captures the primary historical crash periods in China's stock market, such as the 2008 global financial crash, the 2015 market turmoil, the 2018 trade tensions, and the 2020 COVID-19 shock. Interestingly, some periods commonly considered as non-crash cycles are mixed with moments of crash. This should be reasonable given the relatively significant decline in market value. For instance, the average market return was around −10% in January 2002.

3.2 | Macroeconomic and Market-Level Characteristics

In our study, we introduce nine macroeconomic factors. Six of these factors are based on the variable definitions documented in Chen (2009), consisting of unemployment rate (*unemp*), inflation rate (*infl*), M2 money supply growth rate (*m2g*), industrial production growth rate (*ipg*), nominal effective exchange rate (*exr*), and term spread (*tms*). The remaining three factors include gross domestic product growth rate (*gdp*), international trade volume growth rate (*itvg*) and interbank offered rate (*ibor*), which are used in earlier studies as macroeconomic predictors (e.g., Leippold et al. 2022).

In addition to the macroeconomic factors, we also collect a broad set of 24 market-level predictors aggregated from firm-specific and stock-level information to reflect the overall characteristics and future performance of China's stock market, inspired by prior studies on the stock market (e.g., Green et al. 2017; Gu et al. 2020; Leippold et al. 2022).[8] This broad collection of market-level predictors can be primarily grouped into six categories: (i) proxies of stock market trend, such as short-term reversal and market momentum; (ii) proxies of

TABLE 1 | Out-of-sample predictability of evaluated models.

	LOGIT-M	LOGIT	PCA	ENet	LDA	GBCT	SVM	NN
F_1 score	42.62	55.56	62.22	63.41	66.67	73.68	64.86	75.00
G-mean	66.28	67.13	77.14	75.66	78.52	80.32	73.95	82.68
$N\varphi$	63.35	74.22	76.58	77.71	79.57	84.52	79.49	84.91
$N\kappa$	62.17	73.75	76.49	77.70	79.57	84.25	79.10	84.85

Note: This table reports four different monthly out-of-sample evaluation metrics in percentage: (1) F_1 score; (2) G-mean; (3) $N\varphi$; (4) $N\kappa$. The models considered include logistic (LOGIT) regression, LOGIT using only macroeconomic variables (LOGIT-M), principal component analysis (PCA), elastic net (ENet), linear discriminant analysis (LDA), gradient boosted classification tree (GBCT), support vector machine (SVM) and feed-forward neural network with 1 hidden layer (NN). All evaluation metrics are ranged into the interval [0, 1]. Better performed model will lead to larger metric values.

liquidity risk, such as Amihud illiquidity and share turnover; (iii) proxies of stock market volatility, such as stock market variance and return volatility; (iv) proxies of valuation ratios, such as price-to-book and price-to-earnings ratios; (v) proxies of fundamental signals, such as market size and outstanding share; (vi) proxies of systematic risk, such as market beta and beta squared. All market-level characteristics are updated monthly.

As in Gu et al. (2020), our baseline set of predictors also includes the interactions between macroeconomic factors and market-level characteristics. Hence, the total number of covariates is: $9 \times (24 + 1) + 24 = 249$. To avoid the effect of variable scale and look-ahead bias, we standardise all continuous characteristics sample-by-sample with zero mean and unit variance.[9] This is a standard transformation to deal with variables with different scales and has also been used in Boehmer et al. (2008), Coudert and Gex (2008), and Garcia (2013). The Appendix C (Table C.1 and Table C.2) provides the details of these characteristics.[10]

4 | Empirical Analysis

In this section, we investigate the out-of-sample performance of machine learning classifiers in terms of both evaluation metrics and statistical tests, as is commonly done in the literature.

4.1 | Out-of-Sample Performance

Table 1 summarises the out-of-sample predictive accuracy for different machine learning methods in terms of F_1 score, G-mean, $N\varphi$ and $N\kappa$. It is no surprise that all machine learning algorithms outperform the traditional logistic regression models since the ability to deal with high-dimensionality and nonlinearity allows machine learning to be competitive at prediction tasks. When including only nine macroeconomic variables, the LOGIT-M model still has some predictive power in forecasting stock market crashes with F_1 score (42.62%), G-mean (66.28%), $N\varphi$ (63.35%) and $N\kappa$ (62.17%), which indirectly reflects the correlation between the stock market and the overall situation of a country's economy to a certain extent. When including all predictors, all evaluation metrics of the LOGIT model exceed 50% with F_1 score (55.56%), G-mean (67.13%), $N\varphi$ (74.22%) and $N\kappa$ (73.75%), respectively, highlighting the supplementary effect of market-level characteristics to traditional macroeconomic factors in classifying crash/non-crash periods.

When feature selection is utilised, the predictions are significantly improved in terms of all four evaluation metrics. The stronger forecasting power of PCA, ENet, and LDA directly reflects that regularisation and dimension reduction are powerful means to mitigate multicollinearity, avoid overfitting, and improve model performance in high-dimensional settings. Specifically, these three feature selection methods improve the out-of-sample F_1 score to above 60%, with PCA (62.22%), ENet (63.41%) and LDA (66.67%), showing an obvious advantage over the traditional LOGIT and LOGIT-M models. This result provides evidence that some factors are partially redundant and essentially noisy signals in predicting stock market crashes in China. Also, the out-of-sample performance of SVM is comparable to that of feature selection techniques.

The neural network model and the tree-based model, GBCT, further improve the out-of-sample predictability in terms of four evaluation metrics, ranging from 73% to 85%. Such improvement validates the flexibility and superiority of these machine learning tools in learning the complex hidden relations in China's stock market and the interactions between predictors. Overall, being able to accurately approximate complex nonlinear relations, the neural network emerges as the best methodology for predicting stock market recessions with the highest performance measures.

4.2 | Statistical Significance of Model Performance

Using four out-of-sample evaluation metrics for model selection may not be sufficient to fully evaluate the difference in models' performance in reality. To obtain a more comprehensive view of model predictability, we also examine whether the performance differences between models are statistically significant and not merely coincidental. Hence, we first examine the significance of differences among evaluated models by introducing two non-parametric test statistics, including the Friedman test (M. Friedman 1940) and the Iman–Davenport test (Iman and Davenport 1980). These are computed based on the performance metric F_1 score.[11] Specifically, the Friedman test compares the mean ranks of algorithms on multiple data sets (multiple seeds in our study):

$$\chi_F^2 = \left[\frac{12}{N \times K \times (K+1)} \sum_{j=1}^{K} \left(\sum_{i=1}^{N} r_i^j \right)^2 \right] - 3 \times N \times (K+1) \tag{9}$$

with $K-1$ degrees of freedom when $N \geq 10$ and $K \geq 5$ (Demšar 2006) and r_i^j being the rank of the jth of K algorithms on the ith of N random seeds. To avoid an undesirable conservative effect of Friedman's statistic, Iman and Davenport (1980) propose a modified statistic with $K-1$ and $(K-1)(N-1)$ degrees of freedom:

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1) - \chi_F^2} \tag{10}$$

As expected, Table 2 indicates that the statistics of the Friedman and Iman–Davenport tests are 55.34 and 33.98 with p values $< 5\%$, showing a significant rejection of equivalence of the rankings of F_1 score across algorithms at the 5% significance level. It implies that at least one model is different from the others with respect to predictive accuracy. After obtaining the Friedman and the Iman–Davenport test results, we then proceed to conduct a pairwise comparison between two algorithms based on both the parametric paired t-test and the non-parametric Wilcoxon's signed-rank test. Table 3 reports the results of the paired t-test and the Wilcoxon signed-rank test for pairwise comparisons between the models we evaluate. A bold number and asterisk, respectively, denote 5% significance for the paired t-test and the Wilcoxon's signed-rank test. The positive t-test statistic indicates superior predictive accuracy of the column model.

The results reported in Table 3 highlight several interesting facts. The first is that feature selection via PCA, ENet, and LDA produces statistically significant improvements over the LOGIT models. However, there is a slight difference between ENet and LDA. Next, we see statistical differences in the performance between PCA and the other two feature selection methods, providing statistical evidence that there is significant rejection of their performance differences. A possible explanation may be that PCA cannot extract incremental information from the majority of predictors as opposed to ENet and LDA, since it reduces feature dimensions by simply finding a few independent principal components that maximise the variance between predictors. Accordingly, it will more easily overlook some critical time points with a sharp and severe drop in stock market prices. Third, nonlinear methods, including GBCT and SVM, identically improve over linear models, but marginal significance is found in the performance difference between SVM and LDA. The neural network is the only model that produces significant and considerable statistical improvements over linear and nonlinear models with a maximum number of significant pairs.

TABLE 2 | Multiple comparison of monthly out-of-sample prediction.

	Friedman test	Iman–Davenport test
Test statistic	55.34*	33.98*

Note: This table reports the Friedman and Iman–Davenport test statistics comparing differences in the overall performance across models. An asterisk indicates the rejection of the null hypothesis of equivalence across the evaluated models at the 5% significance level.

5 | Investment Decision Implication

Heretofore, our quantitative analysis and statistical tests have validated the predictive superiority of machine learning methods in modelling stock market regimes. We next investigate whether machine learning predictions could better help market participants time market fluctuations and make informed investment decisions. From an economic viewpoint, we consider two types of trading strategies that start with \$1 at the end of December 2011. Each investment decision is guided by the model forecast of the stock market regimes. The first is the long-short strategy: investors hold stocks if the predicted market regime is non-crash and sell short stocks otherwise. Considering the short-selling restrictions in China's stock market, we also include the long-bond strategy: holding stocks in non-crash periods, but investing in risk-free assets otherwise. We use the corresponding 3-month Treasury bill rate as a proxy for risk-free return. It is worthwhile to note that assets are reallocated only when the market regime forecast switches.

5.1 | Full Sample Portfolio

Table 4 reports the performance of value-weighted machine learning portfolios based on all A-share stocks traded at China's

TABLE 3 | Pairwise comparison of monthly out-of-sample prediction.

	LOGIT	PCA	ENet	LDA	GBCT	SVM	NN
LOGIT-M	**8.83***	**7.90***	**14.89***	**14.88***	**12.28***	**12.23***	**29.92***
LOGIT		**2.24***	**5.16***	**6.20***	**6.87***	**5.73***	**16.00***
PCA			1.22*	**2.23***	**3.97***	**2.56***	**7.65***
ENet				1.52	**3.69***	**1.97***	**10.31***
LDA					**2.56***	0.72	**7.46***
GBCT						−1.74	**1.98***
SVM							**4.91***

Note: This table reports the pairwise t-test statistics and the Wilcoxon signed-rank test statistics comparing the out-of-sample predictive performance between two models. Positive numbers indicate the column model outperforms the row model. Bold font indicates the difference is significant at the 5% significance level for pairwise t-test statistics and an asterisk indicates significance at the 5% level for the Wilcoxon signed-rank test.

TABLE 4 | Full-sample portfolio performance.

	BH	LOGIT-M	LOGIT	PCA	ENet	LDA	GBCT	SVM	NN
Long-Short									
Avg (%)	—	1.04	1.94	2.60	3.13	3.27	2.12	3.83	4.97
Cum.R	—	1.18	2.21	2.97	3.57	3.73	2.42	4.37	5.67
SR	—	0.54	1.03	1.39	1.67	1.75	1.12	2.06	2.69
Long-Bond									
Avg (%)	1.17	1.44	1.65	2.11	2.20	2.27	1.85	2.43	2.87
Cum.R	1.34	1.64	1.88	2.41	2.51	2.59	2.11	2.77	3.27
SR	0.62	0.92	0.90	1.30	1.24	1.29	1.13	1.38	1.67

Note: This table reports the out-of-sample portfolio performance measures for all evaluated models of the value-weighted long-short and long-bond portfolios based on the full set of A-share stocks. These measures include the average realised monthly return, cumulative return and annualised Sharpe ratio. All measures are based on realised monthly out-of-sample returns from January 2012 to June 2021.

exchanges during the out-of-sample period. We also report the performance of the buy-and-hold strategy for comparative purposes. The results suggest that machine learning techniques, especially the neural network, are advantageous over traditional logistic regression models by a substantial margin for investment decision support.

First, for the long-short strategy, machine learning models generate cumulative returns ranging between 2.42 and 5.67, all of which are at least 10% higher than those produced by the traditional logistic regression models (LOGIT-M and LOGIT) and 80% higher than that produced by the buy-and-hold strategy. The Sharpe ratios for machine learning long-bond portfolios range between 1.12 and 2.69, exceeding the best traditional model (LOGIT, 1.03) and the buy-and-hold strategy (BH, 0.62). Similarly, for the long-bond strategy, machine learning models also achieve higher economic gains than traditional models.

Second, the neural network model dominates its competitors in both portfolio types and emerges as the best-performing method in terms of economic profitability. For the long-short strategy, the neural network portfolio produces the highest cumulative return (5.67), followed by the SVM (4.37), LDA (3.73), and ENet (3.57), among which ENet and LDA exhibit very similar portfolio performance. This is not surprising, considering the commonalities between these two methods. Correspondingly, the neural network model also has explicit superiority over the other models for long-bond portfolio-level investments.

Interestingly, the long-short strategy of LOGIT-M generates lower economic value than the long-bond strategy. This is potentially due to the trade-off between recall and precision ratios. Also, the GBCT method, which exhibits advantageous out-of-sample predictive performance as shown in the previous section, yields a relatively low increase in cumulative return for both long-short (2.42) and long-bond (2.11) strategies. This finding echoes Leitch and Tanner (1991) and Leippold et al. (2022), who show that there may be a weak connection between statistical significance and economic benefit.

5.2 | Index Portfolio

We also conduct a portfolio analysis based on stock market indexes. There is a primary reason for such practice. At least 80% of the total trading volume on China's stock exchanges is contributed by retail investors with limited funds, for whom it is almost impossible to diversify their investment risks by constructing a full-sample portfolio. Therefore, it seems more practical for them to put money into the index funds that institutional investors manage. In Table 5, we display the out-of-sample cumulative return of machine learning portfolios based on four different types of stock market indexes that effectively differentiate four major segments of China's stock market, including the Shanghai Stock Exchange Composite Index (SSE), Shenzhen Stock Exchange Composite Index (SZSE), China Securities 300 Index (CSI 300) and China Securities Small and Midcap 700 Index (CSI 700). The SSE and SZSE represent the performance of stocks traded on the Shanghai Stock Exchange and the Shenzhen Stock Exchange, respectively. The CSI 300 and CSI 700 are representatives of large-cap and small- and middle-cap stocks traded on China's Stock Exchanges. The neural network model proves robust to this index investment strategy and outperforms all other models considered in this study. In addition, there are two main findings regarding performance heterogeneity in different index portfolios.

On the one hand, the CSI 700 portfolio generates a notably higher economic value than the CSI 300 portfolio under both long-short and long-bond settings. In particular, for the long-short NN portfolio, we get a cumulative return of 5.73 for the CSI 700 index compared with a cumulative return of 4.02 for the CSI 300 index portfolio. This implies that machine learning portfolios based on small- and middle-cap stocks generate a higher cumulative return than those based on large-cap stocks. The observation echoes well with the evidence documented by Næs et al. (2011). Since small firms are relatively more vulnerable to the market crash than their large counterparts, mean-variance investors are more likely to reduce the weight of small-cap stocks in their portfolios during recession periods, making the illiquidity risk of small firms substantially increase. Also, firm size is a proxy for risk. Smaller firms tend to have greater risk

TABLE 5 | Portfolio performance of index investments.

	Cumulative return				Sharpe ratio			
Indexes	**SSE**	**SZSE**	**CSI 300**	**CSI 700**	**SSE**	**SZSE**	**CSI 300**	**CSI 700**
Long-Short								
LOGIT-M	1.28	0.78	2.05	1.06	0.64	0.31	0.95	0.45
LOGIT	0.95	4.03	1.31	2.52	0.48	1.63	0.60	1.08
PCA	1.82	4.20	2.44	3.47	0.92	1.70	1.13	1.50
ENet	2.14	4.88	2.65	3.46	1.08	1.99	1.23	1.50
LDA	2.34	4.70	3.07	4.18	1.19	1.91	1.43	1.82
GBCT	1.31	3.56	1.97	2.44	0.66	1.44	0.91	1.05
SVM	2.43	6.26	3.46	4.36	1.23	2.56	1.62	1.90
NN	3.11	7.96	4.02	5.73	1.59	3.28	1.89	2.51
Long-Bond								
BH	0.63	1.85	1.23	1.09	0.32	0.74	0.56	0.46
LOGIT-M	1.21	1.76	2.01	1.48	0.71	0.90	1.09	0.78
LOGIT	0.87	3.01	1.38	1.86	0.45	1.26	0.66	0.83
PCA	1.36	3.36	2.07	2.43	0.78	1.55	1.09	1.20
ENet	1.42	3.43	2.06	2.29	0.76	1.47	1.02	1.05
LDA	1.51	3.39	2.25	2.57	0.81	1.47	1.12	1.19
GBCT	1.10	3.01	1.79	1.96	0.62	1.38	0.92	0.95
SVM	1.51	3.87	2.36	2.58	0.81	1.66	1.17	1.19
NN	1.78	4.50	2.60	3.07	0.98	1.97	1.31	1.44

Note: This table reports the out-of-sample performance metrics for index portfolios based on different models' predictions, including the cumulative return and the annualised Sharpe ratio. The China's A-share indexes considered include Shanghai Stock Exchange Composite IndeX (SSE), Shenzhen Stock Exchange Composite Index (SZSE), China Securities 300 Index (CSI 300) and China Securities Small and Midcap 700 Index (CSI 700). The measures are based on realised monthly out-of-sample returns from January 2012 to June 2021.

than larger firms due to information uncertainty and limited liquidity, providing investors with higher returns when the market is in an expansion.

On the other hand, portfolios constructed from the SZSE index perform better than portfolios constructed from the SSE index in both long-short and long-bond settings. The difference in stock distribution between the two markets may be a reasonable explanation for such a result. The Shanghai exchange has the bulk of the trading volume and consists of large and state-owned firms, while the Shenzhen exchange consists of small- and medium-sized manufacturing and emerging companies. Similarly, as opposed to large firms, small firms have a slower response to new information and crash events (see, e.g., Lo and MacKinlay 1990; Wang et al. 2009). Thus, small firms appear to have a greater likelihood of immense costs when the stock market is in a recession.

5.3 | Industry Portfolio

We then consider the relative performance of machine learning tools using industry portfolios based on the Guidelines for Industry Classification of Listed Companies issued by the China Securities Regulatory Commission (CSRC) in 2012. Of the 19 industry-sorted portfolios, we exclude five industry portfolios from the analysis because the number of stocks in these industries is fewer than 20. There are two main reasons for such consideration. First, like index portfolios, an industry-sorted portfolio is more practical for retail investors than a full sample portfolio, given the availability of industry investment funds. At the same time, institutional investors more frequently present signals regarding fundamental classifications, such as industries (Choi and Sias 2009). Second, listed companies in different industries face different regulatory and policy environments and have heterogeneous reactions to major emergencies, even within similar macroeconomic conditions (Fiordelisi and Galloppo 2018).

Table 6 displays the cumulative return from investing in each of these 14 industry-sorted portfolios in both long-short and long-only settings. This result demonstrates that almost all machine learning methods potentially improve portfolio performance across different industry sectors. In particular, the neural network model dominates the other machine learning algorithms with the highest cumulative return across almost all industry-sorted portfolios. Moreover, different industry sectors respond to market shocks heterogeneously. The portfolio performances are relatively low for defensive industries involving daily services and necessity goods throughout economic cycles, as well as for

TABLE 6 | Portfolio performance for different industries.

	Long-short								Long-bond								
Models	**LOGIT-M**	**LOGIT**	**PCA**	**ENet**	**LDA**	**GB CT**	**SVM**	**NN**	**BH**	**LOGIT-M**	**LOGIT**	**PCA**	**ENet**	**LDA**	**GB CT**	**SVM**	**NN**
Agriculture	−0.18	1.00	2.65	1.16	1.19	0.50	1.38	2.31	1.00	0.66	1.15	2.10	1.28	1.32	0.95	1.37	1.85
Mining	0.52	−0.26	1.22	0.60	0.92	0.41	0.72	1.19	−0.05	0.38	−0.12	0.61	0.33	0.46	0.26	0.37	0.57
Manufacturing	1.09	3.75	4.42	4.79	5.39	3.73	6.91	8.48	2.01	2.07	3.01	3.62	3.54	3.80	3.23	4.26	4.85
Electric and Heating	0.51	1.26	1.42	2.93	3.10	1.27	2.47	3.38	0.52	0.78	0.94	1.17	1.65	1.71	1.04	1.47	1.80
Construction	1.80	1.08	1.54	2.61	2.74	0.87	2.13	3.16	0.46	1.39	0.83	1.15	1.47	1.52	0.82	1.28	1.67
Wholesale and Retail	0.53	1.44	1.31	2.49	2.80	1.07	2.68	3.85	0.27	0.72	0.87	0.97	1.31	1.42	0.84	1.35	1.73
Transportation	1.84	1.92	2.77	4.31	5.14	1.79	4.35	4.83	0.82	1.67	1.42	1.94	2.37	2.64	1.48	2.36	2.53
Information technology	0.61	4.97	3.66	5.55	4.03	2.43	4.72	8.11	0.98	1.41	2.68	2.67	2.98	2.51	2.09	2.68	3.79
Finance	1.38	0.84	1.64	2.36	1.88	1.35	2.34	3.11	1.16	1.59	1.09	1.63	1.88	1.67	1.43	1.84	2.20
Real estate	0.74	1.12	2.01	2.37	2.43	1.43	2.66	4.83	0.86	1.14	1.11	1.65	1.73	1.77	1.35	1.80	2.60
Business service	1.52	5.31	8.86	7.17	6.51	5.63	7.25	12.27	1.56	2.22	3.36	4.76	4.06	3.88	3.62	4.02	5.51
Scientific research	0.54	5.03	3.12	4.72	4.99	6.17	7.20	9.31	3.08	2.36	4.30	3.71	4.31	4.45	5.06	5.27	6.20
Environment	0.14	5.21	2.35	8.40	6.70	2.49	5.77	8.25	0.31	0.60	2.10	1.44	2.90	2.54	1.45	2.27	2.86
Sports and Entertainment	1.67	2.92	2.40	3.29	2.49	1.49	3.17	4.95	0.31	1.38	1.43	1.41	1.58	1.36	1.03	1.52	2.06

Note: This table reports the cumulative return for industry-based portfolios based on different models' predictions. We consider 14 industry-based portfolios according to the Guidelines for Industry Classification of Listed Companies issued by the China Securities Regulatory Commission (CSRC) in 2012. The industry with fewer than 20 individual stocks are excluded. The measures are based on realised monthly out-of-sample returns from January 2012 to June 2021.

traditional industries dominated by state-owned and large-size enterprises, such as agriculture, mining, electric and heating, and wholesale and retail. A possible interpretation is that these sectors are composed of low beta and low volatility assets, which have relatively stable political connections and minimal vulnerability to market cycles (Novy-Marx 2014). In contrast, aggressive and high-tech sectors characterised by high return volatility achieve a high portfolio performance, such as manufacturing, information technology and science and technology services, given that they are more sensitive to stock market fluctuations, especially during stock market crashes. This behaviour heterogeneity across different industry sectors in China's stock market has also been emphasised by Lee et al. (2013). In general, machine learning methods, especially the neural network model, could better assist market investors in avoiding the inefficient allocation of financial capital between different industry sectors across periods of market expansion and recession.

5.4 | Trading Costs

To further explore the economic significance, we analyse the effect of trading costs on the portfolios' performance. We take into consideration three reasonable levels of transaction costs for decision shift, including 20, 40 and 60 bps. In Table 7, we report the cumulative returns for four index portfolios at three different levels of trading costs.

As expected, despite the infrequency of our trading strategies, the portfolios based on machine learning predictions still achieve an economically significant performance, even after incorporating trading costs. By nature, all cumulative returns decrease with transaction costs. For the NN long-short portfolio based on the SZSE index, the cumulative return decreases from 7.96 to 7.75 under the assumption of 20 bps. In an extreme situation, its cumulative return declines to 7.34. From a more realistic perspective, the analysis of long-bond strategies leads to a similar result. For the SZSE long-bond strategy, the NN cumulative return's reduction is from 4.50 to 4.12 in the case of 60 bps. Overall, these results provide supporting evidence for the remarkable performance of machine learning models, especially the neural network, in generating economic gains after considering trading costs.

6 | Economic Implications

The lack of transparency stands as a widely recognised drawback of machine learning techniques. This opacity of their

TABLE 7 | Index portfolio performance controlling for trading costs.

Trading costs	SSE			SZSE			CSI 300			CSI 700		
	20 bps	40 bps	60 bps	20 bps	40 bps	60 bps	20 bps	40 bps	60 bps	20 bps	40 bps	60 bps
Long-Short												
LOGIT-M	1.18	1.09	1.01	0.71	0.64	0.57	1.93	1.80	1.69	0.97	0.89	0.81
LOGIT	0.90	0.85	0.79	3.90	3.76	3.63	1.25	1.18	1.12	2.42	2.33	2.23
PCA	1.69	1.57	1.44	3.96	3.73	3.50	2.28	2.12	1.98	3.26	3.06	2.87
ENet	2.04	1.95	1.85	4.70	4.52	4.34	2.53	2.42	2.31	3.32	3.18	3.05
LDA	2.24	2.13	2.04	4.52	4.34	4.17	2.94	2.81	2.69	4.02	3.86	3.71
GBCT	1.28	1.24	1.20	3.49	3.42	3.35	1.92	1.88	1.83	2.39	2.33	2.28
SVM	2.35	2.27	2.19	6.09	5.92	5.76	3.36	3.25	3.15	4.23	4.11	3.99
NN	3.01	2.91	2.82	7.75	7.54	7.34	3.90	3.79	3.67	5.57	5.42	5.27
Long-Bond												
BH	0.63	0.62	0.62	1.84	1.84	1.83	1.23	1.22	1.22	1.09	1.08	1.08
LOGIT-M	1.12	1.03	0.95	1.65	1.54	1.44	1.89	1.77	1.66	1.38	1.28	1.19
LOGIT	0.82	0.77	0.72	2.89	2.79	2.68	1.32	1.25	1.19	1.78	1.71	1.63
PCA	1.25	1.15	1.05	3.15	2.96	2.77	1.92	1.79	1.66	2.27	2.12	1.97
ENet	1.35	1.27	1.20	3.29	3.15	3.02	1.97	1.87	1.78	2.19	2.09	1.99
LDA	1.43	1.35	1.28	3.25	3.11	2.98	2.14	2.04	1.95	2.45	2.35	2.24
GBCT	1.07	1.04	1.00	2.95	2.89	2.82	1.74	1.70	1.66	1.91	1.87	1.82
SVM	1.45	1.39	1.34	3.76	3.64	3.53	2.28	2.20	2.13	2.50	2.42	2.33
NN	1.72	1.65	1.59	4.37	4.24	4.12	2.52	2.44	2.35	2.98	2.88	2.79

Note: This table reports the impact of trading costs on cumulative return of the index portfolio for both long-short and long-bond strategies. We consider three types of transaction costs, including 20, 40 and 60 bps. The China's A-share indexes considered include Shanghai Stock Exchange Composite IndeX (SSE), Shenzhen Stock Exchange Composite Index (SZSE), China Securities 300 Index (CSI 300) and China Securities Small and Midcap 700 Index (CSI 700). The measures are based on realised monthly out-of-sample returns from January 2012 to June 2021.

inner workings poses challenges in understanding the decision-making processes of these models, raising concerns about interpretability, accountability and potential biases. In response to this issue, this section endeavours to shed light on the relative importance and influential mechanisms of macroeconomic and market-level characteristics in predicting stock market crashes in China.

6.1 | Feature Importance Ranking

In this subsection, we examine which features drive the predictive performance and investigate existing differences across machine learning models. To quantify variable importance, we employ the SHapley Additive exPlanatory (SHAP) values, which are based on the recent breakthrough in explainable AI by Lundberg and Lee (2017). Precisely, SHAP values can be computed for every combination of input variables and observations. For a given observation, the SHAP value of an input variable represents the extent to which it shifts the expected model output from the unconditional expectation (i.e., the mean of the training data). Hence, the sum of SHAP values for all inputs added to the training sample mean produces the predicted target value. This additive feature also allows us to assess the average importance of a variable across an entire dataset or for a subset of data, as well as measure the aggregate importance of groups of variables.

Individual SHAP values can be either positive or negative. To measure variable importance, we focus on the magnitude of the SHAP value, that is, the absolute value of the SHAP value. For each observation, we scale all absolute SHAP values across the input variables to sum to 100%. These scaled absolute SHAP values ($|SHAP_{j,t}|$) are the relative contribution of an input variable $j \in [1, \ldots J]$ measured at time $t \in [1, \ldots T]$. Therefore, for input variable j across an entire dataset, the variable importance V_j is calculated as:

$$V_j = \frac{1}{T}\sum_{t=1}^{T}\left|SHAP_{j,t}\right| \tag{11}$$

Table 8 shows the results of the feature importance for each macroeconomic variable within each model. The results demonstrate that all methods tend to emphasise a similar set of important macroeconomic variables, but the relative importance of macroeconomic predictors varies across machine learning methods. For example, the international trade volume growth rate (*itvg*) has the largest variable importance for PCA. It is worth noting that *itvg* also has high priority in the other models. PCA also puts substantial weight on *infl* and *m2g*, showing these macroeconomic variables are also influential for PCA. However, *infl* is less important for ENet, LDA and NN. In addition, the distribution of macroeconomic variable importance for ENet, LDA and NN is comparatively more uniform, as opposed to the other methods.

Figure 2 reports the box plot of aggregate macroeconomic variable importance across different machine learning models. Investigating each macroeconomic variable in turn shows that the two most influential variables that stand out in our overall importance rankings across models are term spread (*tms*) and international trade volume growth rate (*itvg*). Meantime, unemployment rate (*unemp*) and M2 money supply growth rate (*m2g*) demonstrate relatively strong explanatory power in predicting China's stock market crashes across models. Other variables, including gross domestic product growth rate (*gdp*), interbank offered rate (*ibor*), and industrial production growth rate (*ipg*), are less important as they are identified as less informative by most models.

TABLE 8 | Variable importance of macroeconomic variables.

	PCA	ENet	LDA	GBCT	SVM	NN
unemp	11.47	13.55	9.36	19.00	13.47	14.78
infl	20.82	4.40	5.05	25.23	11.58	4.97
m2g	17.55	12.51	11.73	13.04	11.43	16.56
ipg	7.34	3.78	4.04	4.12	8.50	4.05
exr	2.78	15.82	15.94	4.77	5.45	8.95
tms	5.96	23.10	21.41	7.43	15.65	22.14
gdp	6.72	6.70	8.74	7.09	7.79	4.38
itvg	22.96	15.64	18.46	12.35	16.30	16.68
ibor	4.40	4.50	5.27	6.97	9.83	7.50

Note: This table reports the variable importance for nine macroeconomic variables in each machine learning model. These factors include: unemployment rate (*unemp*), inflation rate (*infl*), M2 money supply growth rate (*m2g*), industrial production growth rate (*ipg*), nominal effective exchange rate (*exr*), term spread (*tms*), gross domestic product growth rate (*gdp*), international trade volume growth rate (*itvg*) and interbank offered rate (*ibor*). For each model, the sum of variable importances is normalised to one. All values are in percentage.

To discern the varying predictive power of our market-level characteristics in predicting stock market crashes, Figure 3 reports the relative importance rankings of our 24 market-level predictors within each machine learning model. Variable importances within a model are normalised to sum to one, providing the interpretation of relative importance for that specific model. Furthermore, to better understand the importance of groups of market-level characteristics, Figure 4 reports the overall importance rankings of market-level variable groups across machine learning models. Following Geertsema and Lu (2023), we define the feature importance of a variable group for each machine learning method as the sum of the importance of all input variables included in this group.

Figures 3 and 4 show that machine learning models are generally in close agreement about three categories of informative market-level predictors for the prediction of China's stock market crashes. The most influential group comprises momentum-based factors, including short-term reversal (*mom1m*), long-term reversal (*mom36m*) and momentum change (*chmom*). This finding is consistent with the behavioural phenomena prevalent in retail-dominated markets, such as herding and information cascades. In such environments, investors often heavily rely on recent price trends due to limited information or cognitive biases, leading to amplified mispricing and heightened market fragility. Following the momentum group, valuation ratios—such as price-to-cashflow (*pcr*), price-to-book (*pbr*) and price-to-sales (*psr*)—and liquidity-related indicators—such as Amihud illiquidity

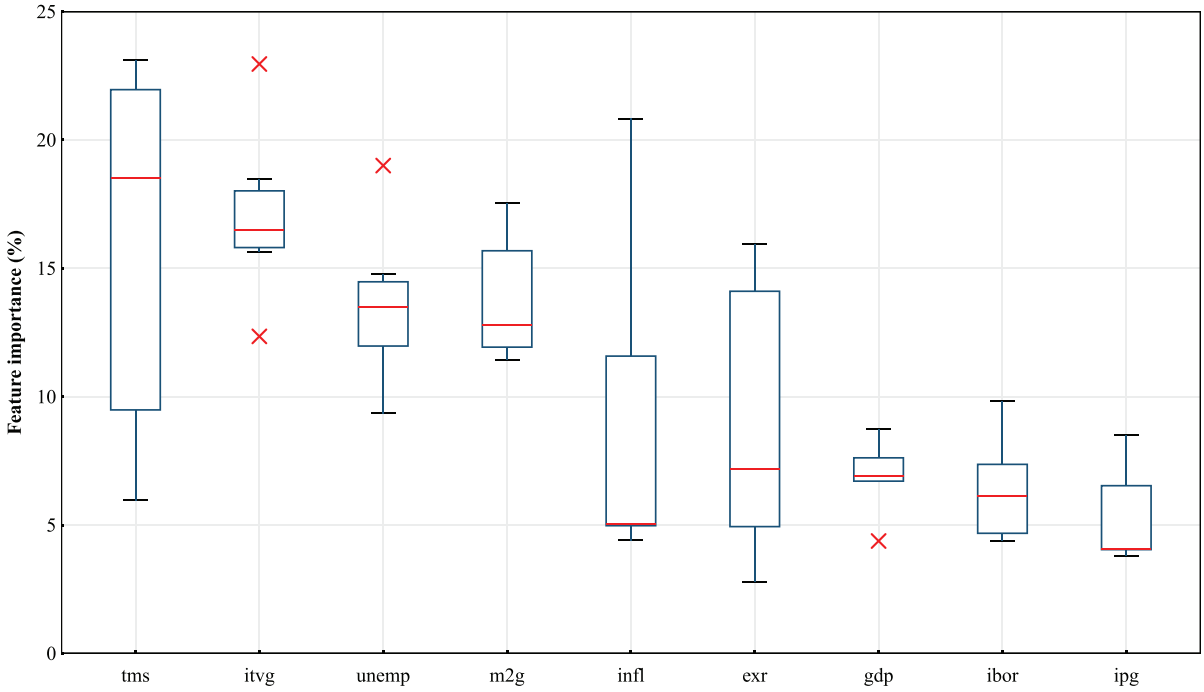

FIGURE 2 | Overall importance for nine macroeconomic variables. This figure displays the aggregate importance of each macroeconomic factor across machine learning models. These factors include: Unemployment rate (*unemp*), inflation rate (*infl*), M2 money supply growth rate (*m2g*), industrial production growth rate (*ipg*), nominal effective exchange rate (*exr*), term spread (*tms*), gross domestic product growth rate (*gdp*), international trade volume growth rate (*itvg*) and interbank offered rate (*ibor*). [Colour figure can be viewed at wileyonlinelibrary.com]

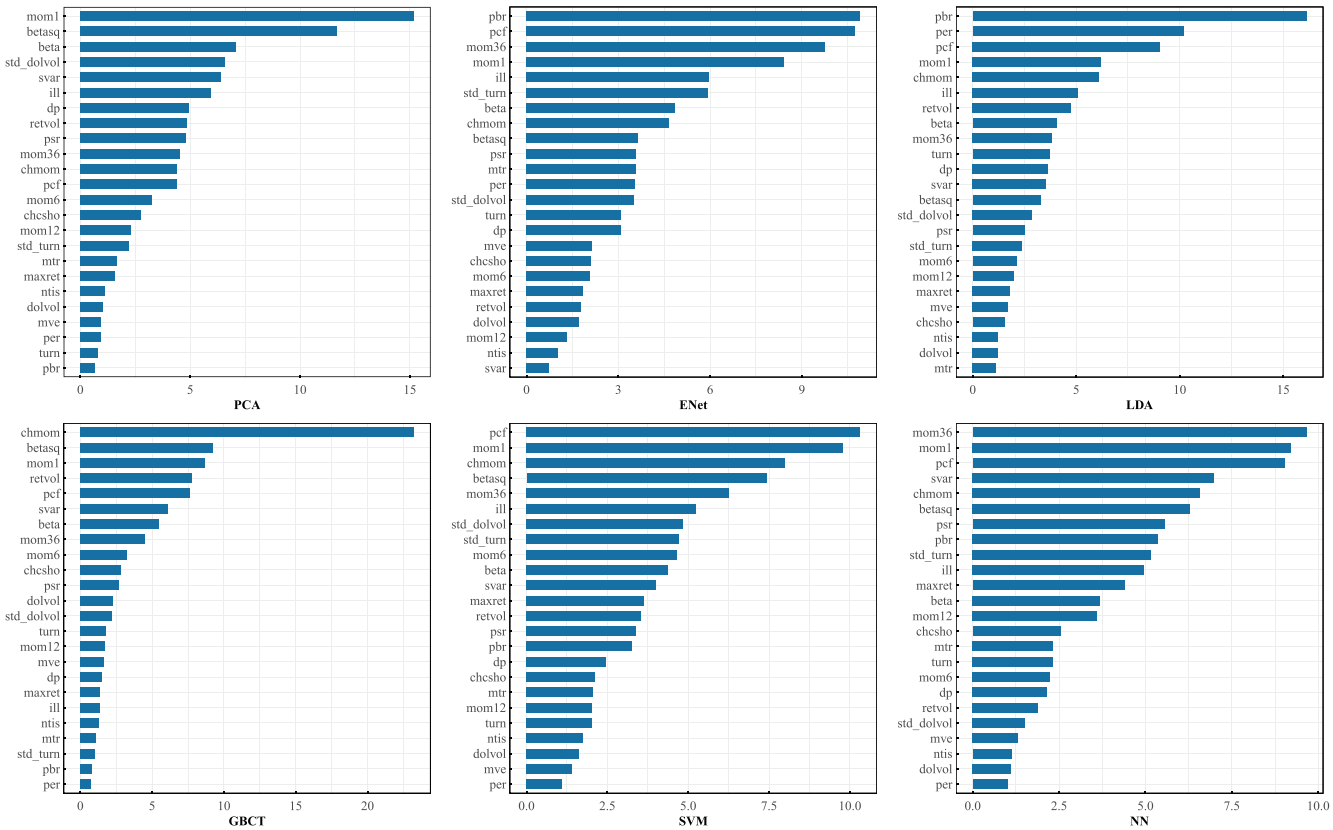

FIGURE 3 | Importance rankings for market-level individual factors. This figure shows the SHAP-based importance for 24 market-level variables. All values are in percentage. [Colour figure can be viewed at wileyonlinelibrary.com]

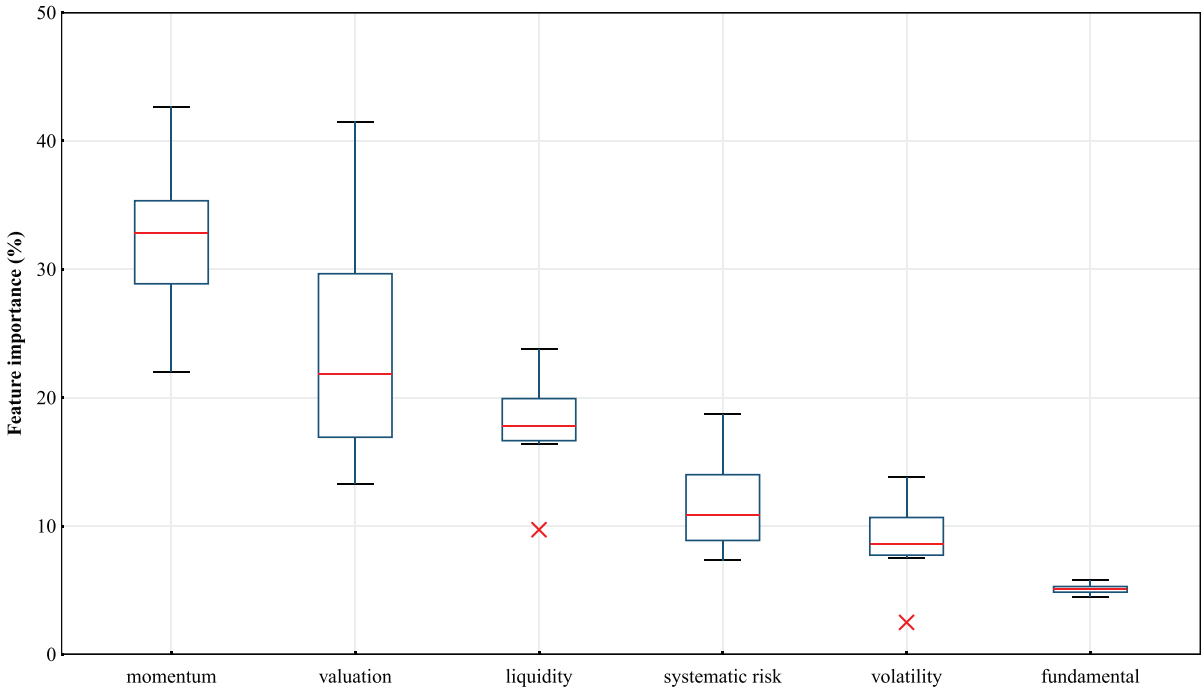

FIGURE 4 | Overall importance for market-level variable groups. This figure displays a box plot for the aggregate variable importance for each of the six market-level variable groups: (1) momentum: *mom1*, *mom6*, *mom12*, *mom36*, *chmom*, *maxret*; (2) liquidity: *ill*, *std_turn*, *std_dolvol*, *dolvol*, *turn*, *mtr*; (3) volatility: *svar*, *retvol*; (4) valuation: *per*, *pbr*, *pcf*, *psr*, *dp*; (5) fundamental: *mve*, *chsco*, *ntis*; (6) systematic risk: *beta*, *betasq*. [Colour figure can be viewed at wileyonlinelibrary.com]

(*ill*) and liquidity volatility (*std_dolvol* and *std_turn*)—also exhibit significant importance. Overall, these patterns are consistent with Gu et al. (2020) and Leippold et al. (2022), who document the strong explanatory power of momentum, liquidity and valuation factors for equity risk premiums.[12]

6.2 | Economic Mechanism

To gain a deeper understanding of the relation between predictor variables and the target response, we elucidate the model-implied marginal effect of individual predictors on the expected probability of stock market crashes. We select our best-performing neural network model for the illustration. In Figure 5, we present the SHAP summary plots for the top five macroeconomic predictors and the top 10 market-level characteristics.[13] Each point in the summary plot corresponds to a SHAP value for a feature and an instance. The *x*-axis depicts the SHAP values of each feature, representing its impact on the model output. A positive (negative) SHAP value indicates that the local feature value contributes to a higher (lower) prediction. The coloration signifies whether the feature has a high (i.e., red) or low (i.e., blue) value for the individual observation.

We unveil three primary observations. First, we show that *tms*, *unemp*, and *infl* are positively associated with SHAP values, indicating that higher values of these features are associated with a greater probability of market crashes. In contrast, *itvg* and *m2g* are negatively related to the probability of stock market crashes. The underlying rationale could be that high unemployment rates, elevated term spreads, and low industrial production growth directly imply an unfavourable economic environment for businesses and investors. Accordingly, the 'flight to safety' sentiment among investors may spread to the equity market, contributing to stock market crashes. Moreover, the positive relation between *infl* and stock market crashes may emphasise the important role of supply-side shocks in China's economy. Additionally, a rise in money supply is related to an investor's portfolio change from non-interest-bearing money to financial assets, including equities, boosting stock market prices. These potential influential mechanisms are consistent with the findings from Chen (2009).

Second, we observe that most market-level characteristics exhibit the expected sign with the target variable (see e.g., Bekaert and Wu 2000; Whitelaw 2000; Li et al. 2005; Coudert and Gex 2008; Andrikopoulos et al. 2014; Fu et al. 2020). Specifically, we demonstrate that higher stock market volatility (i.e., *svar* and *retvol*), overvaluation (i.e., *pcf* and *pbr*), illiquidity risk (i.e., *ill*), and systematic risk (i.e., *beta* and *betasq*) lead to a greater probability of stock market recessions. Moreover, we find that the short-term reversal (i.e., *mom1*) is negatively related to the probability of stock market crashes, while the long-term reversal (i.e., *mom36*) and momentum change (i.e., *chmom*) are positively associated with the probability of stock market crashes. This implies that long-term overreactions and persistent biased beliefs among investors that cannot be justified by fundamentals push up the probability of stock market crashes.

Finally, we reveal that these dominant predictors have a complicated nonlinear effect on the predicted outcome. For example, we observe a dense cluster of instances with low *svar* values

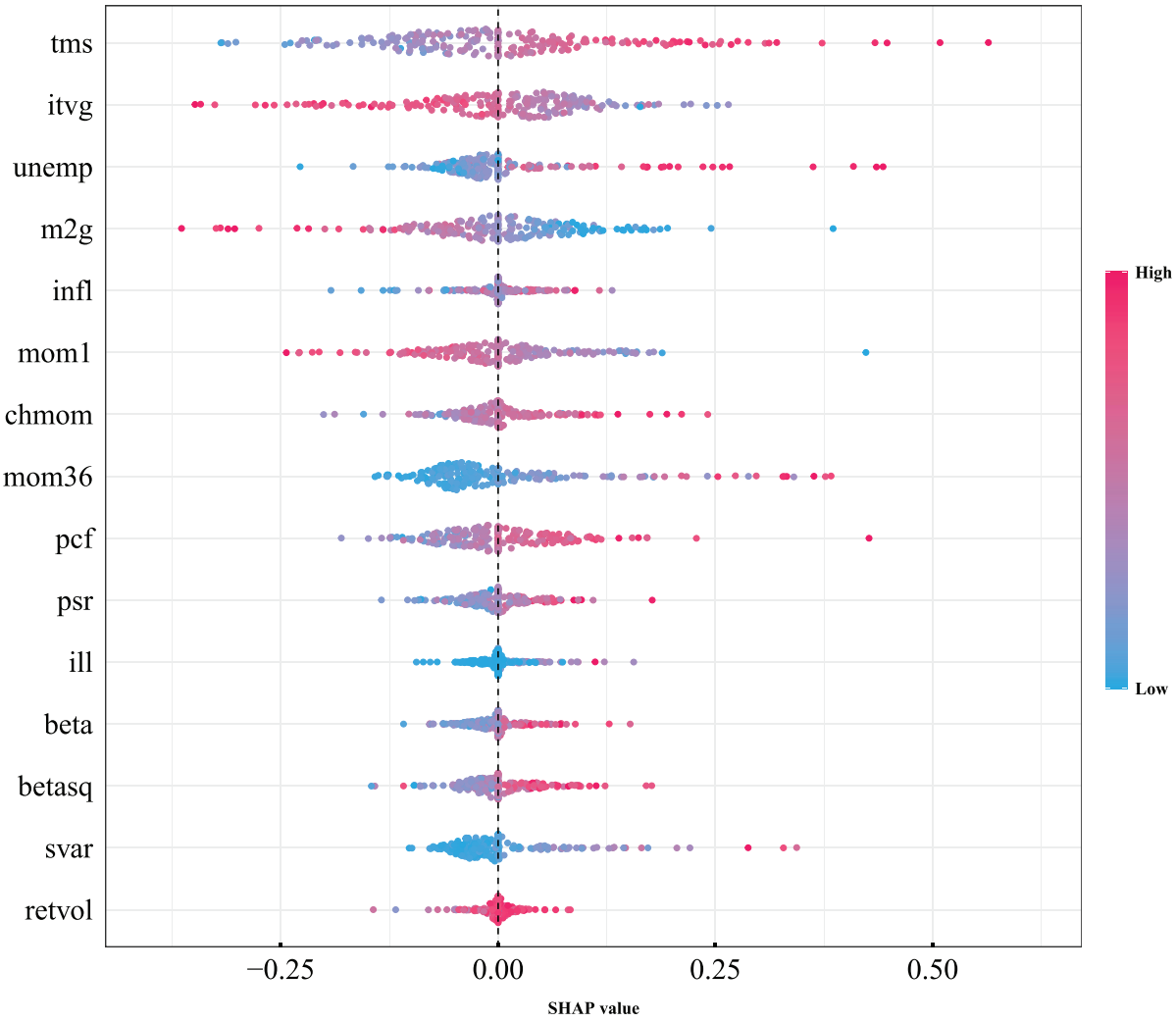

FIGURE 5 | SHAP summary plot. This figure illustrates the SHAP summary plot of the top 5 macroeconomic factors and the top 10 market-level characteristics for the feedforward neural network model. The x-axis represents the SHAP values of each variable and its impact on the model output. The colour reveals whether the feature has a high (i.e., red) or low (i.e., blue) value for the individual observation. [Colour figure can be viewed at wileyonlinelibrary.com]

(blue points) that exhibit small negative SHAP values. As *svar* increases, instances (red points) extend further to the right. This suggests that higher values of *svar* exert a stronger impact on the predicted outcome compared with the lower values of *svar*, signifying the nonlinear impact of *svar* on the expected probability of stock market crashes. A similar phenomenon can be observed for other predictors, such as *unemp* and *mom36*. This intricate relationship further provides a reasonable explanation for the superiority of nonlinear machine learning models (i.e., GBRT and NN) as they have the inherent capability to model the interactions and nonlinearities in the data.

7 | Conclusion

Stock market timing is of primary importance to minimise the social cost and economic loss associated with market crashes and improve the efficiency of financial resource allocation. The degree of predictability rises for stock market crashes should better help these interested parties 'fight fires' ex ante, such as early policy interventions by policymakers and investment decision optimisations by market participants. To this end, we explore the predictive accuracy and economic significance of advanced machine learning techniques in predicting stock market crashes. We aim to develop and filtrate an effective out-of-sample prediction model for China's stock market crashes by comparing a range of machine learning methods.

Our empirical results highlight the potential of machine learning models for modelling stock market fluctuations and improving investment decisions. In particular, the neural network model dominates its competitors and achieves the highest predictive accuracy, whereas traditional econometric models (i.e., logistic regressions) deliver relatively poor predictive performance. For example, using the full set of characteristics, the LOGIT model attains an F_1 score of only 55.56%, compared with 75% for the neural network model. This gap is consistent with the restrictive functional-form and low-dimensionality assumptions of traditional models, which make nonlinear interactions and complex patterns harder to learn. Meanwhile, our portfolio analysis indicates that market timing strategies can largely benefit from the application of machine learning tools. Based

on the portfolio framework, we further confirm large economic gains from the NN model for mean-variance investors through improved out-of-sample predictability. This implies that machine learning predictions could help market participants to improve their investment decisions, be more aware of the risk profile associated with the stock market conditions, and thus track the risk evolution of the stock market over time more precisely. Our results also emphasise the heterogeneity in portfolio performance across different market segments, industry sectors, and stocks. Specifically, stocks with small capitalisations, traded on the Shenzhen Stock Exchange, and within aggressive industry sectors have historically outperformed their corresponding counterparts. This potentially demonstrates their high sensitivity to market crashes and high-risk vulnerability.

Moreover, we demonstrate that aggregate market-level characteristics present considerable supplementary effects to the well-known macro factors. Through an explainable AI technique, we find that nearly all machine learning algorithms produce a very similar ranking of the most explanatory market-level signals. The dominant market-level characteristics are associated with market momentum, valuation, and liquidity. For the future extension of our work, a more sophisticated model with richer information (e.g., unstructured data from media information) might achieve even more accurate predictions of stock market crashes.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Endnotes

[1] It is noteworthy that we apply one hidden layer for the neural network model, following the finding of Gu et al. (2020) that shallow learning achieves better performance than deep learning in asset pricing applications. For a detailed overview of these methods, please see Appendix A.

[2] We require at least 10 years of training data to ensure reliable estimation, as in Bao et al. (2020).

[3] Hyperparameters are tuned by maximising the area under the precision–recall curve (PRAUC). Decision thresholds are then chosen for each classifier by maximising normalised Cohen's kappa ($N\kappa$) over a grid in [0.30,0.60], following Berge and Jordà (2011). Further details on hyperparameter tuning and threshold optimization are provided in Table B.1.

[4] Stratified 3-fold cross-validation preserves the class distribution and reduces the impact of subperiods with few crashes.

[5] See the National Bureau of Statistics of China, the Bank for International Settlements, the Central Depository and Clearing of China, the General Administration of Customs of China, and the People's Bank of China.

[6] The sample period starts in 2002 because part of the data is released from this year, such as the 1-month inter-bank offered rate.

[7] Patel and Sarkar (1998) provide an initial defining criterion of the market crash in both Asian emerging and developed markets, stipulating that it occurs when the equity price index experiences a decline to two standard deviations below the mean of $CMAX_t$. However, the utilisation of this threshold value results in the oversight of significant crash events and yields a restricted number of trigger points within our sample period. To enhance the accuracy of crash identification, we adopt a more conservative approach by setting the threshold value to one standard deviation below the mean of $CMAX_t$. Notably, we also find that the results remain qualitatively identical in unreported tests when the threshold is set to two or three standard deviations below the mean value of $CMAX_t$.

[8] This collection covers nearly all monthly available market-specific signals mentioned in Leippold et al. (2022). It is noteworthy that we exclude consideration of annually and quarterly market-specific variables due to their relatively low importance in conducting monthly prediction tasks related to the stock market, as documented in Gu et al. (2020).

[9] Standardising all data points together will cause look-ahead bias since the mean and variance used are constructed from out-of-sample observations.

[10] Following Gu et al. (2020), we use the most recent monthly characteristics at the end of month t and the most recent quarterly data by the end $t-4$ to predict stock market crashes at month $t+1$ for avoiding another source of the look-ahead bias.

[11] Similar results can be obtained by using the other three evaluation metrics, including G-mean, $N\varphi$, and $N\kappa$.

[12] It is noteworthy that the tree-based model, GBCT, tends to assign less weight to a broad set of predictors than the other methods. Such a difference is potentially due to the generic property of tree-based models. Precisely, they randomly select a subset of market-level features to construct decision trees in the model training process, leading some predictors to become less relevant in some decision trees.

[13] As in Gu et al. (2020), the overall importance ranking of each predictor is determined by the sum of ranks assigned in each method.

References

Amihud, Y. 2002. "Illiquidity and Stock Returns: Cross-Section and Time-Series Effects." *Journal of Financial Markets* 5: 31–56. https://doi.org/10.1016/s1386-4181(01)00024-6.

Andrikopoulos, A., T. Angelidis, and V. Skintzi. 2014. "Illiquidity, Return and Risk in G7 Stock Markets: Interdependencies and Spillovers." *International Review of Financial Analysis* 35: 118–127.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang. 2006. "The Cross-Section of Volatility and Expected Returns." *Journal of Finance* 61: 259–299. https://doi.org/10.1111/j.1540-6261.2006.00836.x.

Bali, T. G., N. Cakici, and R. F. Whitelaw. 2011. "Maxing out: Stocks as Lotteries and the Cross-Section of Expected Returns." *Journal of Financial Economics* 99: 427–446. https://doi.org/10.1016/j.jfineco.2010.08.014.

Banz, R. W. 1981. "The Relationship Between Return and Market Value of Common Stocks." *Journal of Financial Economics* 9: 3–18. https://doi.org/10.1016/0304-405x(81)90018-0.

Bao, Y., B. Ke, B. Li, Y. J. Yu, and J. Zhang. 2020. "Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach." *Journal of Accounting Research* 58: 199–235.

Bekaert, G., and G. Wu. 2000. "Asymmetric Volatility and Risk in Equity Markets." *Review of Financial Studies* 13: 1–42.

Belke, A., and J. Beckmann. 2015. "Monetary Policy and Stock Prices – Cross-Country Evidence From Cointegrated VAR Models." *Journal of Banking & Finance* 54: 254–265.

Berge, T. J., and Ò. Jordà. 2011. "Evaluating the Classification of Economic Activity Into Recessions and Expansions." *American Economic Journal: Macroeconomics* 3: 246–277. https://doi.org/10.1257/mac.3.2.246.

Bjørnland, H. C., and K. Leitemo. 2009. "Identifying the Interdependence Between US Monetary Policy and the Stock Market." *Journal of Monetary Economics* 56: 275–282. https://doi.org/10.1016/j.jmoneco.2008.12.001.

Boehmer, E., C. M. Jones, and X. Zhang. 2008. "Which Shorts Are Informed?" *Journal of Finance* 63: 491–527. https://doi.org/10.1111/j.1540-6261.2008.01324.x.

Boser, B. E., I. M. Guyon, and V. N. Vapnik. 1992. "A Training Algorithm for Optimal Margin Classifiers." In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. IEEE.

Boyd, J. H., J. Hu, and R. Jagannathan. 2005. "The Stock Market's Reaction to Unemployment News: Why Bad News is Usually Good for Stocks." *Journal of Finance* 60: 649–672. https://doi.org/10.1111/j.1540-6261.2005.00742.x.

Bussiere, M., and M. Fratzscher. 2006. "Towards a New Early Warning System of Financial Crises." *Journal of International Money and Finance* 25: 953–973. https://doi.org/10.1016/j.jimonfin.2006.07.007.

Campbell, J. Y., and R. J. Shiller. 1998. "Valuation Ratios and the Long-Run Stock Market Outlook." *Journal of Portfolio Management* 24: 11–26. https://doi.org/10.3905/jpm.24.2.11.

Campbell, J. Y., and S. B. Thompson. 2008. "Predicting Excess Stock Returns out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies* 21: 1509–1531. https://doi.org/10.1093/rfs/hhm055.

Chen, S. S. 2009. "Predicting the Bear Stock Market: Macroeconomic Variables as Leading Indicators." *Journal of Banking & Finance* 33: 211–223. https://doi.org/10.1016/j.jbankfin.2008.07.013.

Chevapatrakul, T. 2013. "Return Sign Forecasts Based on Conditional Risk: Evidence From the UK Stock Market Index." *Journal of Banking & Finance* 37: 2342–2353. https://doi.org/10.1016/j.jbankfin.2013.01.033.

Choi, N., and R. W. Sias. 2009. "Institutional Industry Herding." *Journal of Financial Economics* 94: 469–491. https://doi.org/10.1016/j.jfineco.2008.12.009.

Chordia, T., A. Subrahmanyam, and V. Anshuman. 2001. "Trading Activity and Expected Stock Returns." *Journal of Financial Economics* 59: 3–32. https://doi.org/10.1016/s0304-405x(00)00080-5.

Cortes, C., and V. Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20: 273–297. https://doi.org/10.1007/bf00994018.

Coudert, V., and M. Gex. 2008. "Does Risk Aversion Drive Financial Crises? Testing the Predictive Power of Empirical Indicators." *Journal of Empirical Finance* 15: 167–184. https://doi.org/10.1016/j.jempfin.2007.06.001.

Datar, V. T., Y. Naik, and R. Radcliffe. 1998. "Liquidity and Stock Returns: An Alternative Test." *Journal of Financial Markets* 1: 203–219. https://doi.org/10.1016/s1386-4181(97)00004-9.

Demšar, J. 2006. "Statistical Comparisons of Classifiers Over Multiple Data Sets." *Journal of Machine Learning Research* 7: 1–30.

Dhatt, M. S., Y. H. Kim, and S. Mukherji. 1999. "The Value Premium for Small-Capitalization Stocks." *Financial Analysts Journal* 55: 60–68. https://doi.org/10.2469/faj.v55.n5.2300.

Dichtl, H., W. Drobetz, and T. Otto. 2023. "Forecasting Stock Market Crashes via Machine Learning." *Journal of Financial Stability* 65: 101099.

Duda, R. O., P. E. Hart, and D. G. Stork. 2000. *Pattern Classification*. Wiley-Interscience.

Estrella, A., and F. S. Mishkin. 1998. "Predicting US Recessions: Financial Variables as Leading Indicators." *Review of Economics and Statistics* 80: 45–61. https://doi.org/10.1162/003465398557320.

Fama, E. F., and J. D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81: 607–636. https://doi.org/10.1086/260061.

Ferson, W. E., and C. R. Harvey. 1997. "Fundamental Determinants of National Equity Market Returns: A Perspective on Conditional Asset Pricing." *Journal of Banking & Finance* 21: 1625–1665.

Fiordelisi, F., and G. Galloppo. 2018. "Stock Market Reaction to Policy Interventions." *European Journal of Finance* 24: 1817–1834. https://doi.org/10.1080/1351847x.2018.1450278.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29: 1189–1232.

Friedman, M. 1940. "A Comparison of Alternative Tests of Significance for the Problem of *m* Rankings." *Annals of Mathematical Statistics* 11: 86–92. https://doi.org/10.1214/aoms/1177731944.

Fu, J., Q. Zhou, Y. Liu, and X. Wu. 2020. "Predicting Stock Market Crises Using Daily Stock Market Valuation and Investor Sentiment Indicators." *North American Journal of Economics and Finance* 51: 100905.

Garcia, D. 2013. "Sentiment During Recessions." *Journal of Finance* 68: 1267–1300. https://doi.org/10.1111/jofi.12027.

Geertsema, P., and H. Lu. 2023. "Relative Valuation With Machine Learning." *Journal of Accounting Research* 61: 329–376.

Genkin, A., D. D. Lewis, and D. Madigan. 2007. "Large-Scale Bayesian Logistic Regression for Text Categorization." *Technometrics* 49: 291–304. https://doi.org/10.1198/004017007000000245.

Gettleman, E., and J. M. Marks. 2006. *Acceleration Strategies. Working Paper.* University of Illinois at Urbana-Champaign.

Green, J., J. R. Hand, and X. F. Zhang. 2017. "The Characteristics That Provide Independent Information About Average US Monthly Stock Returns." *Review of Financial Studies* 30: 4389–4436. https://doi.org/10.1093/rfs/hhx019.

Gu, S., B. Kelly, and D. Xiu. 2020. "Empirical Asset Pricing via Machine Learning." *Review of Financial Studies* 33: 2223–2273. https://doi.org/10.1093/rfs/hhaa009.

Harvey, C. R. 1988. "The Real Term Structure and Consumption Growth." *Journal of Financial Economics* 22: 305–333. https://doi.org/10.1016/0304-405x(88)90073-6.

Hau, H., and H. Rey. 2005. "Exchange Rates, Equity Prices, and Capital Flows." *Review of Financial Studies* 19: 273–317. https://doi.org/10.1093/rfs/hhj008.

Iman, R. L., and J. M. Davenport. 1980. "Approximations of the Critical Region of the Fbietkan Statistic." *Communications in Statistics - Theory and Methods* 9: 571–595. https://doi.org/10.1080/03610928008827904.

Jegadeesh, N. 1990. "Evidence of Predictable Behaviour of Security Returns." *Journal of Finance* 45: 881–898. https://doi.org/10.1111/j.1540-6261.1990.tb05110.x.

Jegadeesh, N., and S. Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *Journal of Finance* 48: 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x.

Khandani, A. E., A. J. Kim, and A. W. Lo. 2010. "Consumer Credit-Risk Models via Machine-Learning Algorithms." *Journal of Banking & Finance* 34: 2767–2787.

Kingma, D. P., and J. Ba. 2014. "Adam: A Method for Stochastic Optimization." *arXiv Preprint*. https://doi.org/10.48550/ARXIV.1412.6980.

Kuvshinov, D., and K. Zimmermann. 2022. "The Big Bang: Stock Market Capitalization in the Long Run." *Journal of Financial Economics* 145: 527–552. https://doi.org/10.1016/j.jfineco.2021.09.008.

Lee, C. C., M. P. Chen, and K. M. Hsieh. 2013. "Industry Herding and Market States: Evidence From Chinese Stock Markets." *Quantitative Finance* 13: 1091–1113. https://doi.org/10.1080/14697688.2012.740571.

Leippold, M., Q. Wang, and W. Zhou. 2022. "Machine Learning in the Chinese Stock Market." *Journal of Financial Economics* 145: 64–82. https://doi.org/10.1016/j.jfineco.2021.08.017.

Leitch, G., and J. E. Tanner. 1991. "Economic Forecast Evaluation: Profits Versus the Conventional Error Measures." *American Economic Review* 81, no. 3: 580–590.

Leroy, A., and A. Pop. 2019. "Macro-Financial Linkages: The Role of the Institutional Framework." *Journal of International Money and Finance* 92: 75–97. https://doi.org/10.1016/j.jimonfin.2018.12.002.

Li, C., S. R. Tan, N. Ho, and W. M. Chia. 2022. "Behavioral Heterogeneity and Financial Crisis: The Role of Sentiment." *Physica A: Statistical Mechanics and Its Applications* 603: 127767.

Li, Q., J. Yang, C. Hsiao, and Y. J. Chang. 2005. "The Relationship Between Stock Returns and Volatility in International Stock Markets." *Journal of Empirical Finance* 12: 650–665.

Li, T., S. Zhu, and M. Ogihara. 2006. "Using Discriminant Analysis for Multi-Class Classification: An Experimental Investigation." *Knowledge and Information Systems* 10: 453–472. https://doi.org/10.1007/s10115-006-0013-y.

Lo, A. W., and A. C. MacKinlay. 1990. "When Are Contrarian Profits due to Stock Market Overreaction?" *Review of Financial Studies* 3: 175–205.

Lu, S., C. Liu, and Z. Chen. 2021. "Predicting Stock Market Crisis via Market Indicators and Mixed Frequency Investor Sentiments." *Expert Systems With Applications* 186: 115844. https://doi.org/10.1016/j.eswa.2021.115844.

Lundberg, S. M., and S. I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, et al., 4765–4774. Curran Associates, Inc.

Morck, R., B. Yeung, and W. Yu. 2000. "The Information Content of Stock Markets: Why Do Emerging Markets Have Synchronous Stock Price Movements?" *Journal of Financial Economics* 58: 215–260. https://doi.org/10.1016/s0304-405x(00)00071-4.

Næs, R., J. A. Skjeltorp, and B. A. Ødegaard. 2011. "Stock Market Liquidity and the Business Cycle." *Journal of Finance* 66: 139–176. https://doi.org/10.1111/j.1540-6261.2010.01628.x.

Nagel, S. 2005. "Short Sales, Institutional Investors and the Cross-Section of Stock Returns." *Journal of Financial Economics* 78: 277–309. https://doi.org/10.1016/j.jfineco.2004.08.008.

Novy-Marx, R. 2014. *Understanding Defensive Equity. Working Paper.* University of Rochester. https://doi.org/10.3386/w20591.

Nyberg, H. 2013. "Predicting Bear and Bull Stock Markets With Dynamic Binary Time Series Models." *Journal of Banking & Finance* 37: 3351–3363. https://doi.org/10.1016/j.jbankfin.2013.05.008.

Patel, S. A., and A. Sarkar. 1998. "Crises in Developed and Emerging Stock Markets." *Financial Analysts Journal* 54: 50–61. https://doi.org/10.2469/faj.v54.n6.2225.

Paye, B. S. 2012. "'déjà vol': Predictive Regressions for Aggregate Stock Market Volatility Using Macroeconomic Variables." *Journal of Financial Economics* 106: 527–546. https://doi.org/10.1016/j.jfineco.2012.06.005.

Pontiff, J., and A. Woodgate. 2008. "Share Issuance and Cross-Sectional Returns." *Journal of Finance* 63: 921–945. https://doi.org/10.1111/j.1540-6261.2008.01335.x.

Rapach, D. E., J. K. Strauss, and G. Zhou. 2013. "International Stock Return Predictability: What is the Role of the United States?" *Journal of Finance* 68: 1633–1662. https://doi.org/10.1111/jofi.12041.

Roll, R. 1988. "R2." *Journal of Finance* 43: 541–566. https://doi.org/10.1111/j.1540-6261.1988.tb04591.x.

Tay, F. E., and L. Cao. 2001. "Application of Support Vector Machines in Financial Time Series Forecasting." *Omega* 29: 309–317. https://doi.org/10.1016/s0305-0483(01)00026-3.

Thorbecke, W. 1997. "On Stock Market Returns and Monetary Policy." *Journal of Finance* 52: 635–654. https://doi.org/10.1111/j.1540-6261.1997.tb04816.x.

Tu, J. 2010. "Is Regime Switching in Stock Returns Important in Portfolio Decisions?" *Management Science* 56: 1198–1215. https://doi.org/10.1287/mnsc.1100.1181.

Wang, J., G. Meric, Z. Liu, and I. Meric. 2009. "Stock Market Crashes, Firm Characteristics, and Stock Returns." *Journal of Banking & Finance* 33: 1563–1574. https://doi.org/10.1016/j.jbankfin.2009.03.002.

Welch, I., and A. Goyal. 2007. "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction." *Review of Financial Studies* 21: 1455–1508. https://doi.org/10.1093/rfs/hhm014.

Whitelaw, R. F. 2000. "Stock Market Risk and Return: An Equilibrium Approach." *Review of Financial Studies* 13: 521–547.

Zhang, R., X. Xian, and H. Fang. 2019. "The Early-Warning System of Stock Market Crises With Investor Sentiment: Evidence From China." *International Journal of Finance and Economics* 24: 361–369.

Zou, H., and T. Hastie. 2005. "Regularisation and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 67: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Appendix A

Details of Machine Learning Models

This section provides detailed descriptions of all models used in this study, consisting of regularised logistic regression, principal component analysis (PCA), linear discriminant analysis (LDA), gradient boosting classification tree (GBCT), support vector machine (SVM) and feedforward neural network (NN).

Regularised Logistic Regression

Logistic regression is an extensively employed model to make predictions of stock market movements and crashes in literature. Specifically, it quantifies the probability of stock market recession as a function of a predictor set:

$$P(S_t = 1|\boldsymbol{X}) = \frac{e^{\boldsymbol{\beta X}}}{1 + e^{\boldsymbol{\beta X}}} \tag{A.1}$$

where $P(S_t = 1|\boldsymbol{X})$ is the conditional probability of stock market crash given the information set of predictors. $\boldsymbol{\beta}$ is a M-dimensional vector of regression coefficients, which are estimated via minimising the following objective function:

$$\mathcal{L}_H(\boldsymbol{\beta}) = \frac{1}{T}\sum_{t=1}^{T} H\left(S_t, h_\beta(x_t)\right) \tag{A.2}$$

where the cross-entropy loss $H(\cdot)$ is a function of the stock market regime and the predicted conditional probability of stock market crash:

$$H\left(S_t, h_\beta(x_t)\right) = \begin{cases} -\log\left(h_\beta(x_t)\right) & \text{if } S_t = 1 \\ -\log\left(1 - h_\beta(x_t)\right) & \text{if } S_t = 0 \end{cases} \tag{A.3}$$

However, the logistic regression with a lack of shrinkage tends to overfit the data in case of a large covariate set. Especially, in high-dimensional settings, it leads to low model ability to generalise new data due to the inconsistency of coefficient estimates caused by multicollinearity (Genkin et al. 2007). To avoid overfitting and improve model stability, one of the most popular solutions is the penalised regression approach which constrains the number of regression coefficients via automatically deleting or shrinking the effect of the unnecessary covariates. Specifically, it imposes a regularisation term on the objective function of a logistic regression model, resulting in the coefficients of less informative factors toward or equal to zero. The most widely used penalty terms in generalised linear regression models consist of LASSO, Ridge and Elastic net (ENet) corresponding to different regularisation effects. The ENet combines characteristics of both LASSO and Ridge penalties. It shrinks some coefficient estimates to be exactly zero (like LASSO) and some coefficient estimates close to zero (like Ridge). The objective function for ENet is specified as:

$$L_H^{\text{ENet}}(\boldsymbol{\beta}; \lambda, \rho) = L_H(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{M}\left[(1-\rho)\left|\beta_j\right| + \frac{1}{2}\rho\beta_j^2\right] \tag{A.4}$$

where λ regulates the extent of shrinkage, and ρ is the ENet mixing parameter. The ENet is a general case of LASSO ($\rho = 0$) and Ridge ($\rho = 1$). In our application, we focus on the role of elastic net penalty. We also refer readers to Zou and Hastie (2005) for more detailed descriptions of elastic net.

Principal Component Analysis

In contrast with regularisation methods that classify objects directly in high-dimensional space, principal component analysis (PCA) is an unsupervised dimensionality-reduction tool to reduce data dimension first, and then classify objects in a new feature space. Specifically, it reduces the number of predictors in the initial data set by generating new uncorrelated variables that are linear combinations of initial features. Hence, PCA transforms initial data from a high-dimensional space into a lower-dimensional space in such a way that the linearly projected data contains as much as possible of variance. For the implementation of PCA, there are three main procedures. The first step is to calculate the covariance matrix of all initial variables. The second step is to calculate the eigenvector and corresponding eigenvalues of the covariance matrix to identify the principal components. The final step is to use part of the leading components in a standard logistic regression model. In particular, the logistic regression with PCA can be specified in the following form:

$$O = \left(XW_K\right)\beta_K + \tilde{\varepsilon} \tag{A.5}$$

where O is a $T \times 1$ vector of log odds of conditional probability of $P(S_{t+1} = 1|\boldsymbol{X})$, X is the corresponding $T \times M$ matrix of stacked covariates, W_K is a $M \times K$ orthogonal transformation matrix, β_K is the $K \times 1$ vector of principal component parameters, and $\tilde{\varepsilon}$ is the $T \times 1$ vector of model residuals.

Linear Discriminant Analysis

Another popular dimension reduction technique for tackling classification problems is linear discriminant analysis (LDA), which may be considered a simple application of Bayes' Theorem for classification. LDA is like principal component analysis, but it focuses on maximizing the separability among known classes. Specifically, in the Bayes framework, the conditional probability is given by:

$$P(S_t = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell} \tag{A.6}$$

where π_k is the prior probability of class k and $f_k(x)$ is the class-conditional density of X in class S_t. It should be mentioned that LDA assumes each class density follows multivariate Gaussian distribution and the classes have an identical covariance matrix Σ. However, LDA frequently achieves good performances in classification tasks, even if the assumptions of normality and common covariance matrix among classes are often violated (Li et al. 2006; Duda et al. 2000). Then, in comparing two classes k and ℓ, the log-odd ratio generated by LDA is formulated by:

$$\log\frac{P(S_t = k|X = x)}{P(S_t = \ell|X = x)} = \beta_0 + x\beta_1 \tag{A.7}$$

where $\beta_0 \equiv \log\frac{\pi_k}{\pi_\ell} - \frac{1}{2}\left(\mu_k + \mu_\ell\right)^T\Sigma^{-1}\left(\mu_k - \mu_\ell\right)$, and $\beta_1 \equiv \Sigma^{-1}\left(\mu_k - \mu_\ell\right)$. According to (A.7), the decision function of LDA is given by:

$$\Psi_k^{\text{LDA}}(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k \tag{A.8}$$

However, when the number of training samples is small compared with the number of features, the estimated covariance matrix becomes highly unstable. One of the possible solutions in the literature is to introduce a shrinkage parameter to adjust the estimation of the class covariance matrix:

$$\Sigma_{\text{shrunk}} = (1-\lambda)\hat{\Sigma} + \lambda\frac{\text{Tr}\hat{\Sigma}}{p}\text{Id} \tag{A.9}$$

In general, selecting a successful shrinkage parameter is an important task. A detailed description of shrinkage tuning is presented in Table B.1.

Gradient Boosting Classification Tree

Gradient boosting classification tree is a powerful ensemble-learning tool for classification analysis. The basic idea of this method is to sequentially apply multiple 'weak learners' to produce a 'strong learner'.

In detail, it starts by computing initial predictions through building a simple tree:

$$T_0(x,\Theta)=\sum_{j=1}^{J} c_j I\left(x\in R_j\right) \tag{A.10}$$

where $R_1, \ldots, R_J$ represent a partition of the feature space, $I(\cdot)$ is an indicator function, $\Theta=\left(c_j,R_j\right)_{j=1}^{J}$ are the unknown parameters. Then, it fits a second tree the residuals from the initial predictions. Hence, at each interaction $b=1, \ldots, B$, the boosting model can be described as:

$$F_b(x)=F_{b-1}(x)+\xi^{*}\sum_{j=1}^{J} c_{jb} I\left(x\in R_{jb}\right) \tag{A.11}$$

where $\xi\,(0<\xi\le 1)$ is the shrinkage factor, also called learning rate, that is used to scale the contribution of each base tree model for preventing over-fitting and improving model performance. A lower learning rate has the effect of making fewer corrections for each tree. This in turn results in more trees that should be added to the model. In addition, the unknown parameters can be estimated as:

$$\min_{\Theta_b}\sum_{i=1}^{N} L\left(y_i,F_b\left(x_i\right)\right) \tag{A.12}$$

where N is the number of instances, and $L(.)$ is a loss function. We also refer readers to J. H. Friedman (2001) for a detailed description of this method.

In our application, we consider several tuning parameters to modify the gradient boosting tree algorithm, namely, the total number of trees, learning rate, the maximum depth of the tree and a minimum number of observations falling in each leaf.

Support Vector Machine

Support vector machine is a long-established method for classifying both linearly and nonlinearly separable data. The method has great potential to handle high-dimensional data and improve the overfitting issue in learning tasks since the number of covariates in training data does not have a distinct influence on model complexity. This property would also be beneficial for crash identification because there may be no limit to the number of predictors on which crash forecasting is based. In detail, let us start with a fundamental formulation of the so-called two-class support vector classification.

For linearly separable data, the SVM method constructs a hyperplane in feature space. Here, a hyperplane is a subspace of dimensionality $D-1$, where D is the number of dimensions of the feature space itself. For example, in the two-dimensional feature space of Figure A.1 below, the one-dimensional hyperplane ($\boldsymbol{w}\cdot\boldsymbol{x}+b=0$) illustrated in red separates the 'blue' class from the 'green' class. Model optimisation is performed by finding a maximum-margin decision boundary for the hyperplane, by using so-called support vectors (data points on $\boldsymbol{w}\cdot\boldsymbol{x}+b=\pm 1$):

$$\begin{aligned}&\min_{w,b}\frac{1}{2}\|w\|^2\\ &\text{s.t. } y_i\left(w\cdot x_i+b\right)\ge 1;\qquad i=1,2,\cdots,N\end{aligned} \tag{A.13}$$

where $\boldsymbol{w}$ is an adjustable weight vector, and b is a bias term. However, in practice, there are few cases in which training data are perfectly linearly separable due to the existence of noise and outliers. Hence, the slack variable will be introduced to permit misclassification:

$$\begin{aligned}&\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i\\ &\text{s.t. }\quad y_i\left(w\cdot x_i+b\right)\ge 1-\xi_i;\quad \xi_i\ge 0,\qquad i=1,2,\cdots,N\end{aligned} \tag{A.14}$$

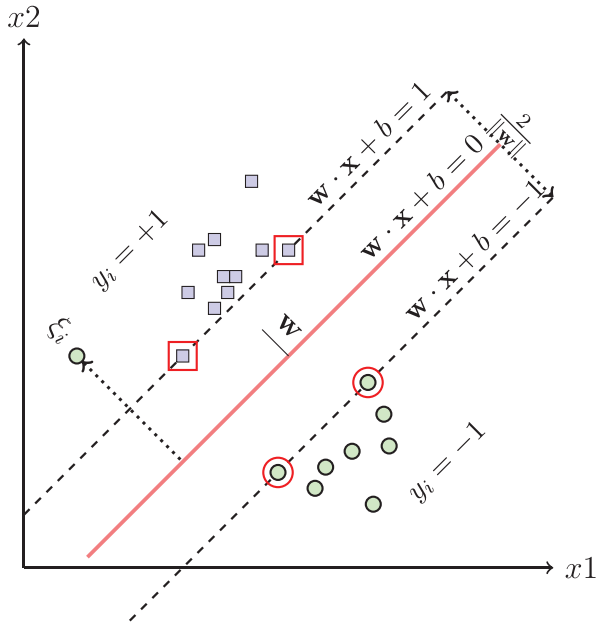

FIGURE A.1 | Illustration of linearly separable data. [Colour figure can be viewed at wileyonlinelibrary.com]

where ξ is the slack variable which is used for avoiding the violation of (A.13); C is a regularisation constant, which is introduced to adjust the penalty intensity for misclassified points. If the penalty parameter is large, SVM will choose a small margin to minimise classification errors and vice versa.

Furthermore, for nonlinearly separable data, an illustration is displayed in Figure A.2 for intuitively understanding the implementation of the nonlinear SVM. As shown in Figure A.2, the binary data points are nonlinearly separable in a feature space x. The SVM method could map the input data points into a high-dimensional feature space z by using the nonlinear mapping, φ, associated with a kernel function. As a result, the data points that are not linearly separable in original x space, could be separated by a linear hyperplane with the help of a kernel function. Specifically, we have multiple selections for the kernel function:

$$K\left(x,x_k\right)=\begin{cases}x_k^T x, & \text{Linear}\\ \left(x_k^T x+\theta\right)^p, & \text{Polynomial}\\ \exp\left\{-\gamma\|x-x_k\|^2\right\}, & \text{Radial Basis Function}\\ \tanh\left(\gamma x_k^T x+\theta\right), & \text{Sigmoid}\end{cases} \tag{A.15}$$

Then, similar to linearly separable data, the optimisation procedure for nonlinearly separable data will be:

$$\begin{aligned}&\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i\\ &\text{s.t. } y_i\left(w\varphi\left(x_i\right)+b\right)\ge 1-\xi_i;\quad \xi_i\ge 0,\qquad i=1,2,\cdots,N\end{aligned} \tag{A.16}$$

For further details, we refer readers to the classical textbooks and literature (Cortes and Vapnik 1995; Boser et al. 1992). In this study, the radial basis function is used as the kernel function of SVM due to high empirical performance in classification tasks. We consider two hyperparameters since they play a decisive role in the performance of SVM as evidenced by Tay and Cao (2001). One is the regularisation constant mentioned above. The other is the kernel coefficient in the radial basis kernel function, which is used to adjust the distance impact of a single training point. For high values of the kernel coefficient, fewer points are

grouped together and the regions separating different classes get more specific, tending to overfit the training data and vice versa.

Feed-Forward Neural Network

The final machine learning method analysed in this study is the feed-forward neural network, which uses hidden layers with a large number of joined neurons and nonlinear transformations to capture the complex nonlinear relations between features. In recent years, the neural network has gained considerable popularity in both practice and academia as the growing amount of computational power and data. In particular, neural networks have been applied to handle many real-life issues, such as vehicle control, medical diagnosis and handwritten text identification.

In our implementation, we focus on feed-forward neural networks with one hidden layer that contains m neurons, with the idea of achieving better performance about shallow learning documented in Gu et al. (2020). As shown in Figure A.3, the basic architecture of neural network, like the human brain, usually consists of three types of classes: the input layer, which brings the initial information into the system for further analysis; the hidden layer, where the complex relationship between features in the input data is identified; the output layer, which produces the output result of given input data from the hidden layer. The Figure A.4 provides a detailed illustration of signal flow at a specific neuron. A neuron extracts information linearly from all input units $x_n^{(1)}$. Then, it transforms the aggregate signal from these inputs into derived feature $O_m^{(2)}$ through applying a 'activation function' (φ), which is the source of non-linearity. Specifically, a neuron transforms its inputs into an output as $O_m^{(2)} = \varphi\left(b_m^{(1)} + \sum w_{mn}^{(1)} x_n^{(1)}\right)$. Finally, the output layer linearly aggregates the derived features from each neuron $\left(b^{(2)} + \sum w_m^{(2)} O_m^{(2)}\right)$ into the final prediction.

For the activation function used to transform the input data, we choose the hyperbolic tangent function (Tan$H$), commonly used in binary classification practice. The function is given as:

$$\text{Tan}H(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{A.17}$$

To estimate the parameters in a neural network, we adopt the adaptive moment estimation (Adam) by Kingma and Ba (2014) for computational efficiency. Other tuning parameters are presented in Table B.1.

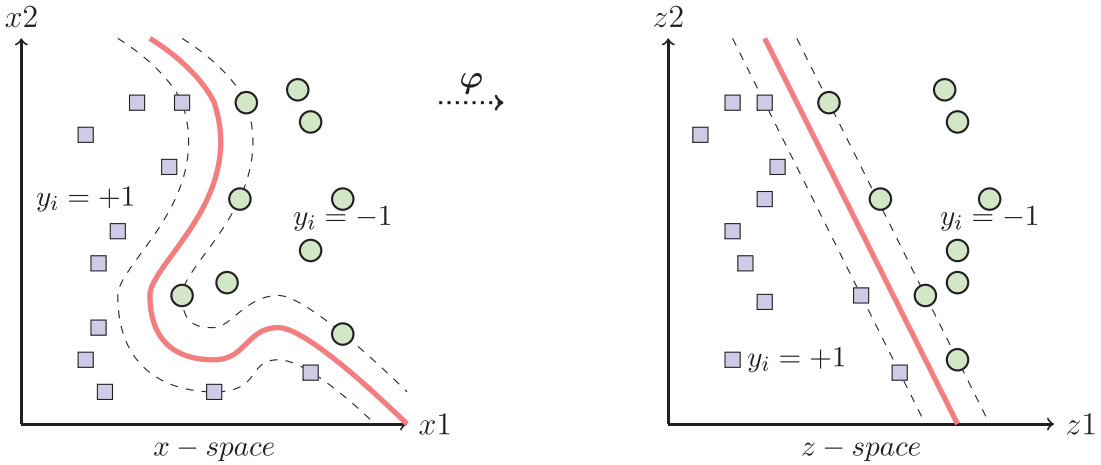

FIGURE A.2 | Illustration of nonlinearly separable data. [Colour figure can be viewed at wileyonlinelibrary.com]

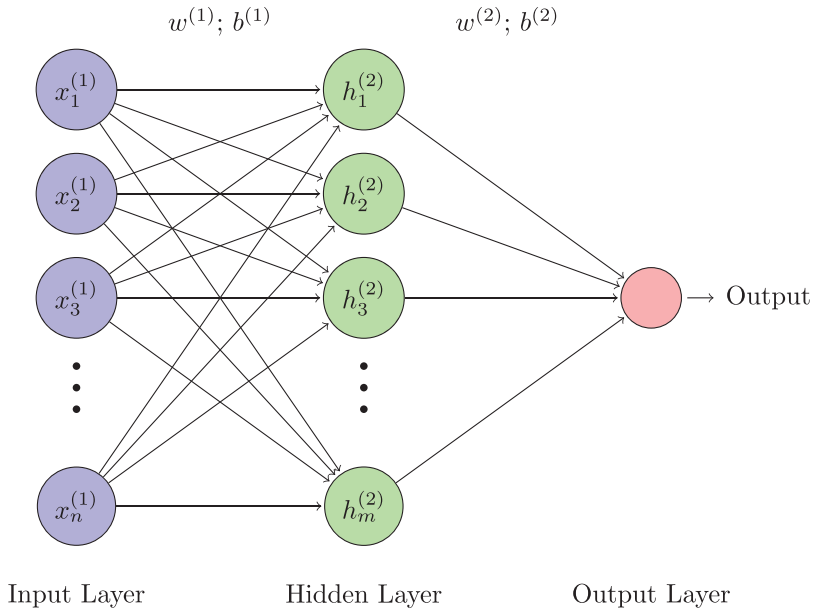

FIGURE A.3 | Illustration of feed-forward neural network. [Colour figure can be viewed at wileyonlinelibrary.com]

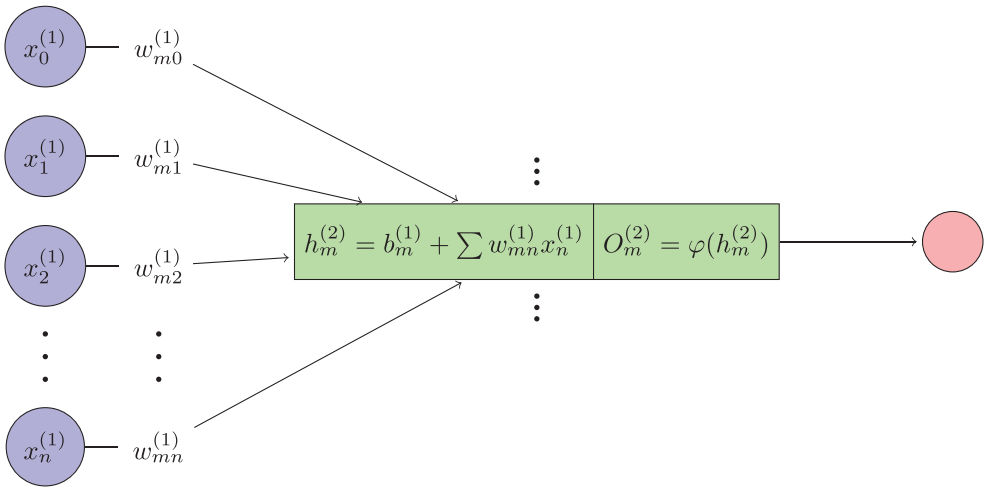

FIGURE A.4 | Illustration of signal flow at neuron m. [Colour figure can be viewed at wileyonlinelibrary.com]

Appendix B

Hyperparameters

TABLE B.1 | Listing of hyperparameters for each machine learning model.

	PCA	ENet	LDA	GBCT	SVM	NN
Ensemble = 10 Threshold $t = 0.3–0.6$	√	√	√	√	√	√
Others	K	$\lambda \in$ $(10^{-4}, 10^{-1})$	$\lambda \in$ $(10^{-4}, 10^{-1})$	Depth $\in \{3, 6, 9\}$ Trees $\in \{50, 100, 200\}$ Leaf $\in \{5, 10, 15\}$ LR $\in \{0.01, 0.1\}$	$C \in (10^1, 10^4)$ $\gamma \in (10^{-4}, 10^{-1})$	Neurons $\in \{8, 16, 24, 32\}$ $\lambda \in (10^{-5}, 10^{-2})$ LR $\in \{0.001, 0.01\}$

Appendix C

The List of Characteristics

This section describes a broad set of characteristics that are used in our study, consisting of 24 market-level characteristics and 9 macroeconomic variables.

Market-Level Characteristics

Table C.2 provides a detailed description of the 24 market-level characteristics for the prediction of China's stock market crashes.

TABLE C.1 | The list of market-level characteristics.

No.	Acronym	Definition	Frequency	Source
1	amihud	Average of daily (absolute return/RMB trading volume).	Monthly	Amihud (2002)
2	beta	Average of stock-level beta with 24 months of returns.	Monthly	Fama and MacBeth (1973)
3	betasq	Beta squared.	Monthly	Fama and MacBeth (1973)
4	chcsho	Percent change in market share outstanding.	Monthly	Pontiff and Woodgate (2008)
5	chmom	Cumulative returns from months $t-6$ to $t-1$ minus months $t-12$ to $t-7$.	Monthly	Gettleman and Marks (2006)
6	dolvol	Natural logarithm of average daily RMB trading volume from month $t-2$.	Monthly	Chordia et al. (2001)
7	dp	Natural logarithm of (12-month moving sum of dividends of all A-share stocks/weighted stock price).	Monthly	Welch and Goyal (2007)
8	maxret	Maximum daily return from returns during month $t-1$.	Monthly	Bali et al. (2011)
9	mom1	Cumulative return of month $t-1$.	Monthly	Jegadeesh and Titman (1993)
10	mom6	Cumulative return from months $t-6$ to $t-2$.	Monthly	Jegadeesh and Titman (1993)
11	mom12	Cumulative return from months $t-12$ to $t-2$.	Monthly	Jegadeesh (1990)
12	mom36	Cumulative return from months $t-36$ to $t-13$.	Monthly	Jegadeesh and Titman (1993)
13	mtr	Ratio of average daily RMB trading volume to the average daily market value.	Monthly	Welch and Goyal (2007)
14	mve	Natural log of market capitalisation at end of month $t-1$.	Monthly	Banz (1981)
15	ntis	12-month moving sums of market net issues divided by the total market capitalisation.	Monthly	Welch and Goyal (2007)
16	pbr	Market capitalisation divided by book value of all A-share stocks.	Monthly	Nagel (2005)
17	pcr	Natural logarithm of (weighted stock price/12-month moving sum of cash flow of all A-share stocks).	Monthly	Ferson and Harvey (1997)
18	per	Natural logarithm of (weighted stock price/12-month moving sum of earnings of all A-share stocks).	Monthly	Campbell and Shiller (1998)
19	psr	Natural logarithm of (weighted stock price/12-month moving sum of sales of all A-share stocks).	Monthly	Dhatt et al. (1999)
20	retvol	Standard deviation of daily returns from month $t-1$.	Monthly	Ang et al. (2006)
21	std_dolvol	Standard deviation of daily RMB trading volume.	Monthly	Chordia et al. (2001)
22	std_turn	Standard deviation of daily share turnover.	Monthly	Chordia et al. (2001)
23	svar	Sum of squared daily returns.	Monthly	Welch and Goyal (2007)
24	turn	Average daily trading volume from month $t-3$ to $t-1$ scaled by the number of shares outstanding at end of month t.	Monthly	Datar et al. (1998)

Macroeconomic Characteristics

TABLE C.2 | The list of macroeconomic characteristics.

No.	Acronym	Macroeconomic factor	Frequency	Source
1	unemp	The change in unemployment rate (YoY) from the National Bureau of Statistics of China.	Quarterly	Boyd et al. (2005); Chen (2009)
2	infl	The change in consumer price index (YoY) from the National Bureau of Statistics of China.	Monthly	Welch and Goyal (2007); Chen (2009)
3	m2g	The M2 money supply growth rate (YoY) from the National Bureau of Statistics of China.	Monthly	Welch and Goyal (2007); Chen (2009)
4	ipg	The industrial production growth rate (YoY) from the National Bureau of Statistics of China.	Monthly	Chen (2009)
5	exr	The nominal exchange rate index from Bank for International Settlements (BIS).	Monthly	Hau and Rey (2005); Chen (2009b)
6	tms	The spread between the yield on 3-month government bond and the 10-year government bond from China Central Depository and Clearing.	Monthly	Estrella and Mishkin (1998); Chen (2009)
7	gdp	The gross domestic product growth rate (YoY) from the National Bureau of Statistics of China.	Quarterly	Paye (2012); Kuvshinov and Zimmermann (2022)
8	itvg	The international trading volume growth rate (YoY) from the General Administration of Customs of China.	Monthly	Rapach et al. (2013); Leippold et al. (2022)
9	ibor	The 1-month inter-bank offered rate from The People's Bank of China.	Monthly	Belke and Beckmann (2015)