Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Illusion of Accuracy: Machine-Learning and GARCH Forecasts of UK Equity Index Prices

Uzair Bhatti

uzair.bhatti@ukmc.ac.uk

UK Management College

Sadaat Umer

University of Dundee

Awais Iqbal

UK Management College

Research Article

Keywords: Stock price prediction, Machine learning, GARCH, Forecast evaluation, Volatility forecasting, FTSE

Posted Date: October 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11480555/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The Illusion of Accuracy: Machine-Learning and GARCH Forecasts of UK Equity Index Prices

Abstract

This study examines whether artificial intelligence (AI) models predict UK stock index prices more accurately than simple benchmarks. Using weekly FTSE 100 and FTSE 250 data from January 2015 to October 2026, it compares eight models in an expanding-window design with 251 out of sample weeks, the random walk, the historical mean, an autoregressive integrated moving average (ARIMA) model, autoregressive models with symmetric and asymmetric (GJR) generalised autoregressive conditional heteroscedasticity (GARCH) errors, a predictive regression, a random forest and a neural network. Forecasts are evaluated on price errors, formal tests of return predictability, forecast-bias regressions, volatility and Value-at-Risk backtests, and trading returns after costs. All models produce predicted prices close to actual prices, with mean absolute percentage errors of 1.3–1.6%, but none beats the random walk significantly, and the random forest and predictive regression are significantly less accurate for the FTSE 100. No forecast-based strategy outperforms buy-and-hold for the FTSE 100 after transaction costs. In contrast, the GJR-GARCH model provides the most accurate volatility forecasts, significantly better than the RiskMetrics benchmark for the FTSE 250, and correctly calibrated Value-at-Risk for both indices. The results reveal an illusion of accuracy in comparisons of predicted and actual prices and indicate that the value of quantitative models in UK equity markets lies in measuring risk rather than predicting returns.

Keywords: Stock price prediction; Machine learning; GARCH; Forecast evaluation; Volatility forecasting; FTSE

JEL classification: C45; C53; C58; G14; G17

1 Introduction

Artificial intelligence (AI) is increasingly presented as a tool that can predict stock prices. Machine-learning models are now standard in academic forecasting research (Gu et al. 2020; Kumbure et al. 2022), and generative AI has put forecasting tools in the hands of retail investors, whose use of these tools measurably affects trading volume (Cheng et al. 2025). Much of the evidence offered for AI forecasting, however, takes the form of charts in which predicted prices lie close to actual prices, or of low percentage errors in price levels. Such evidence is weak. Because stock prices are highly persistent, a forecast that simply repeats last week's price will also lie close to the actual price. The relevant question is not whether AI predictions look close to actual prices, but whether they are closer than a naive benchmark and whether the difference is statistically and economically meaningful.

The question matters for two reasons. First, the Efficient Market Hypothesis implies that no model using past prices should consistently beat the random walk (Fama 1970), and decades of out of sample evidence suggest that apparent predictability is fragile (Welch and Goyal 2008; McLean and Pontiff 2016). If AI models nonetheless succeed, that would be evidence of exploitable inefficiency; if they fail, investors relying on AI forecasts may be misled. Second, while machine learning dominates recent research on price prediction, GARCH-family models remain the standard tool for forecasting risk (Bollerslev 1986; Glosten

et al. 1993; Hansen and Lunde 2005). Few studies compare the two approaches directly on both the level and the risk of prices within one design, and evidence for the UK market is particularly scarce, UK-affiliated authors produce only around 5% of research on AI stock prediction (Lin et al. 2024).

This study addresses that gap. It compares predicted and actual prices for the FTSE 100 and FTSE 250 using weekly data from January 2015 to October 2026. Eight models are evaluated: two benchmarks (the random walk and historical mean), three econometric models (ARIMA, AR–GARCH and AR–GJR-GARCH with Student-t errors) and three learning models (a predictive regression, a random forest and a neural network). Each model is re-estimated weekly in an expanding window and evaluated over 251 out-of-sample weeks covering the 2022 interest-rate shock and the 2025 tariff sell-off. The study asks three research questions:

1. **RQ1.** Do machine-learning models predict UK index prices and returns more accurately than the random walk and econometric benchmarks?
2. **RQ2.** Do GARCH-family models forecast the volatility and downside risk of UK index returns more accurately than historical benchmarks?
3. **RQ3.** Do forecast differences translate into economic value for investors after transaction costs?

The results show that every model's predicted prices track actual prices closely, with mean absolute percentage errors of about 1.3–1.6%, yet no model beats the random walk significantly. The random forest and predictive regression are significantly less accurate than the random walk for the FTSE 100, and no forecast-based trading strategy outperforms buy-and-hold in the rising FTSE 100 market. In contrast, the asymmetric GJR-GARCH model provides the most accurate volatility forecasts and reliable Value-at-Risk estimates for both indices.

The paper makes three contributions. First, it provides recent out-of-sample evidence for the UK, extending a literature dominated by US data. Second, it evaluates AI and GARCH models side by side on four criteria: price accuracy, return predictability, volatility forecasting and economic value. Third, it demonstrates an “illusion of accuracy” in comparisons of predicted and actual prices and sets out a simple evaluation standard for judging claims about AI forecasting tools.

The remainder of the paper is organised as follows. Section 2 reviews the literature and develops the hypotheses. Section 3 describes the data and methodology. Section 4 presents the findings, Section 5 discusses them and Section 6 concludes.

2 Literature review and hypotheses

2.1 Market efficiency and the random-walk benchmark

The Efficient Market Hypothesis holds that prices reflect available information, so past prices should not predict future returns (Fama 1970). Its practical implication for forecasting is that the random walk, which predicts that next week's price equals this week's price, is the benchmark any model must beat. Out-of-sample studies of return predictability show how hard that is: predictors that work in-sample routinely fail against the historical mean once evaluated in real time (Welch and Goyal 2008), and even when they work, the gains are small and depend on economically sensible restrictions (Campbell and Thompson 2008). Predictability also decays once it is published and traded on, with returns on documented anomalies falling by more than half after publication (McLean and Pontiff 2016). The Adaptive Markets

Hypothesis reconciles these findings by treating efficiency as a changing condition shaped by competition among strategies (Lo 2004).

2.2 Econometric models of returns and volatility

The Box–Jenkins ARIMA framework models the conditional mean of a series through its own lags and past shocks (Box and Jenkins 1970). Stock returns, however, show little linear dependence in the mean but strong dependence in the variance: large changes cluster together. The ARCH model (Engle 1982) and its generalisation GARCH (Bollerslev 1986) capture this volatility clustering, and the Student-t version accommodates the fat tails of returns (Bollerslev 1987). Because bad news tends to raise volatility more than good news of the same size, the GJR-GARCH model adds an asymmetric term (Glosten et al. 1993), reflecting the leverage effect first documented by Black (1976). A large comparison of 330 volatility models found little evidence that any beats a simple GARCH(1,1) for exchange rates, although asymmetric models did better for equities (Hansen and Lunde 2005). For forecasting, the key distinction is that GARCH-type models predict the *size* of future moves well but say little about their *direction* (Poon and Granger 2003).

2.3 Machine learning and artificial intelligence in price prediction

Machine-learning methods relax the linearity of ARIMA by learning non-linear interactions among many inputs. Random forests average many decision trees grown on resampled data, which makes them robust to noise and overfitting (Breiman 2001), and feed-forward neural networks can approximate any continuous function given enough hidden units (Hornik et al. 1989). In finance, neural networks and tree ensembles roughly doubled the performance of regression-based return forecasts across US stocks (Gu et al. 2020), and long short-term memory networks outperformed memory-free classifiers on S&P 500 constituents (Fischer and Krauss 2018). Reviews of the field find these methods dominate recent research but are usually judged by error metrics on short samples rather than by economic value (Kumbure et al. 2022; Lin and Lobo Marques 2024).

The most recent wave uses large language models and pre-trained time-series foundation models. Language models reading news headlines predict next-day returns, but strategy returns decline as their adoption rises (Lopez-Lira and Tang 2023), and part of their reported accuracy reflects memorised outcomes rather than prediction (Lopez-Lira et al. 2025; Gao et al. 2026). Off-the-shelf time-series foundation models underperform in zero-shot return forecasting (Rahimikia et al. 2025), and the advantages reported for language-model trading agents deteriorate over longer and broader samples (Li et al. 2025).

2.4 Evidence on economic value and the research gap

Where economic value is tested, machine-learning gains shrink. Return signals lose much of their profitability once microcaps are excluded and trading costs are charged (Avramov et al. 2023), and simple linear models outperform several transformer forecasters on standard benchmarks (Zeng et al. 2023). Three gaps follow. First, most head-to-head comparisons of AI and econometric models use US data, with UK evidence rare (Lin et al. 2024). Second, AI models are seldom compared directly with GARCH-family models on both price and volatility forecasts within the same design. Third, few studies test whether forecast gains survive formal significance tests and transaction costs.

2.5 Hypotheses

The review motivates five testable hypotheses.

- **H1.** Machine-learning models produce more accurate one-week-ahead price forecasts than the random walk.
- **H2.** Machine-learning models produce more accurate return forecasts than ARIMA and GARCH-family models.
- **H3.** Forecast accuracy relative to the random walk is higher for mid-capitalisation stocks (FTSE 250) than for large-capitalisation stocks (FTSE 100).
- **H4.** GARCH-family models, and the asymmetric GJR-GARCH in particular, forecast weekly volatility more accurately than historical-volatility benchmarks.
- **H5.** Model-based forecasts generate trading returns that exceed buy-and-hold after transaction costs.

3 Data and methodology

3.1 Data and sample

The sample covers the two main UK equity indices: the FTSE 100, comprising the largest companies listed on the London Stock Exchange, and the FTSE 250, comprising the next 250 by market capitalisation. Weekly closing, opening, high and low prices and trading volume were obtained from Investing.com for 4 January 2015 to 18 October 2026, giving 616 weekly prices and 615 weekly returns per index. One FTSE 250 volume observation (week of 25 August 2024) was missing and was set equal to the previous week's value; no price data were missing. Weekly frequency was chosen because it filters day-of-the-week noise while providing enough observations for GARCH estimation and out-of-sample testing; monthly data would leave too few test observations for reliable inference.

The sample is divided into a training period from January 2015 to December 2021 (364 weekly returns), which includes the 2016 Brexit referendum and the 2020 COVID-19 crash, and an out-of-sample test period from 2 January 2022 to 18 October 2026 (251 weeks). The test period covers the 2022 interest-rate shock, the 2023 regional banking stress, the April 2025 tariff sell-off and the period in which generative AI tools became widely available to investors. Weekly log returns in per cent are computed as

$$R_t = \ln(P_t / P_{t-1}) \qquad \textit{(Equation 1)}$$

where Pt is the closing index level in week t and Rt is a log return for period t.

3.2 Forecasting models

Eight models generate one-week-ahead return forecasts. Each return forecast is converted into a price forecast as $\hat{P}(t+1) = P(t) \times \exp(\hat{r}(t+1)/100)$, so that predicted and actual prices can be compared directly.

Model	Type	Specification
Random walk (RW)	Benchmark	Next week's price equals this week's price ($\hat{r} = 0$)
Historical mean (HM)	Benchmark	Mean of all past weekly returns

Model	Type	Specification
ARIMA	Econometric	ARMA(p, q) on returns, equivalent to ARIMA(p, 1, q) on log prices; order chosen by the Bayesian information criterion (BIC) over p, q ≤ 2 in the training sample, excluding the white-noise case already covered by the historical mean
AR–GARCH	Econometric	AR(1) mean with GARCH(1,1) variance and Student-t errors
AR–GJR-GARCH	Econometric	AR(1) mean with asymmetric GJR-GARCH(1,1) variance and Student-t errors
Predictive regression (OLS)	Statistical learning	Linear regression of next week's return on the 12 predictors below
Random forest (RF)	Machine learning (AI)	300 trees, minimum 10 observations per leaf, half the predictors considered at each split
Neural network (MLP)	Machine learning (AI)	Feed-forward network with two hidden layers (16 and 8 neurons), L2 penalty, early stopping; average of five networks with different random starts

The GJR-GARCH model, the most general econometric specification, is

$$r_t = \mu + \phi(r_{t-1} - \mu) + \varepsilon_t, \qquad \varepsilon_t = \sqrt{h_t} z_t, \qquad z_t \sim t_\nu \qquad \textit{(Equation 2)}$$

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \gamma \varepsilon_{t-1}^2 I(\varepsilon_{t-1} < 0) + \beta h_{t-1} \qquad \textit{(Equation 3)}$$

where ht is the conditional variance, I(·) equals one after a negative shock, and z_t follows a standardised Student-t distribution with ν degrees of freedom. Setting γ = 0 gives the symmetric GARCH(1,1). Parameters are estimated by maximum likelihood, and persistence is measured as α + β + γ/2.

The three learning models use the same 12 predictors, all known at the end of week t: the current and three lagged weekly returns; the absolute return; 4- and 12-week rolling standard deviations; 4-, 12- and 26-week momentum (cumulative returns); the weekly high–low range; the log change in trading volume; and the current return of the other index.

3.3 Out-of-sample design

Forecasts are produced with an expanding window. For each of the 251 test weeks, every model is re-estimated using only data available up to the previous week, and a forecast is made for the following week. This walk-forward design reproduces the information set of an investor in real time and prevents look-ahead bias. The ARMA order is fixed from the training sample, while all parameters are re-estimated weekly.

3.4 Evaluation of price and return forecasts

Price accuracy is measured by the root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) between predicted and actual index levels. Because price errors are dominated by the level of the index, the main tests are run on returns. The out-of-sample R^2 of Campbell and Thompson (2008) compares each model with the historical mean:

$$R^2_{OS} = 1 - \frac{\sum_t (r_t - \hat{r}_t)^2}{\sum_t (r_t - \bar{r}_t)^2} \qquad \textit{(Equation 4)}$$

where r̄(t) is the historical-mean forecast; a positive value means the model beats the mean.

Four tests assess whether differences are statistically meaningful. The Diebold and Mariano (1995) test, with the small-sample correction of Harvey et al. (1997), compares squared forecast errors against the random walk. Because the random walk is nested within the other models, the Clark and West (2007) test is also reported. The Pesaran and Timmermann (1992) test assesses whether the predicted direction of price changes is correct more often than chance. Finally, the Mincer and Zarnowitz (1969) regression of actual on predicted returns,

$$r_{t+1} = \alpha + b\hat{r}_{t+1} + u_{t+1} \qquad \textit{(Equation 5)}$$

tests forecast unbiasedness through the joint hypothesis a = 0 and b = 1, using Newey and West (1987) standard errors.

3.5 Evaluation of volatility forecasts

Weekly volatility is unobserved, so it is proxied by the range-based estimator of Parkinson (1980), which uses the weekly high and low and is far less noisy than the squared return:

$$\sigma^2_{PK,t} = \frac{[100(\ln H_t - \ln L_t)]^2}{4 \ln 2} \qquad \textit{(Equation 6)}$$

GARCH and GJR-GARCH variance forecasts are compared with two historical benchmarks: the exponentially weighted moving average used by RiskMetrics with a decay factor of 0.94 (J.P. Morgan/Reuters 1996) and a 12-week rolling variance. Accuracy is measured by the mean squared error and the QLIKE loss, which is robust to noise in the volatility proxy (Patton 2011). Value-at-Risk (VaR) forecasts at the 95% and 99% levels are derived from the Student-t GARCH models and backtested with the unconditional coverage test of Kupiec (1995).

3.6 Economic value

To test whether forecasts have economic value, a simple strategy holds the index in week t+1 when the predicted return is positive and holds cash otherwise. A transaction cost of 0.10% is charged each time the position changes, and the cash return is set to zero for comparability across models. Each strategy is compared with buy-and-hold using the annualised return, annualised volatility, Sharpe ratio and maximum drawdown.

3.7 Software and use of artificial intelligence tools

All estimation was carried out in Python 3.13 using NumPy, SciPy, pandas and scikit-learn. The ARMA, GARCH, GJR-GARCH and test procedures were coded directly and validated on simulated data with known

parameters before use. In line with the journal's policy, a large language model (Claude, Anthropic) was used to assist in writing the analysis code and drafting the manuscript; the author checked the code, results and text and takes full responsibility for the content.

4 Findings and analysis

4.1 Properties of UK index returns

Fig. 1 shows the two indices over the full sample. The FTSE 100 rose from about 6,500 to 10,675, with most of the gain after 2024, while the FTSE 250 fell sharply in 2022 and only regained its early-2022 level in 2026.

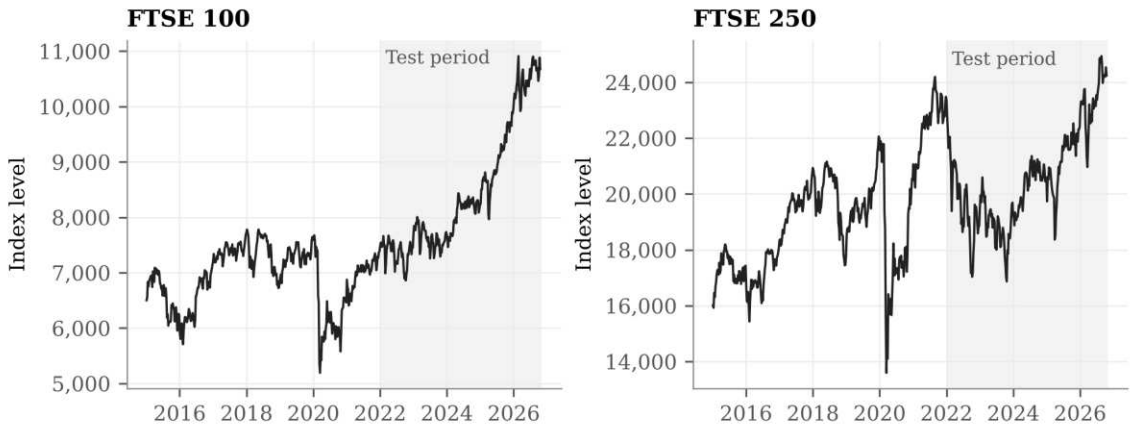


Fig. 1 Weekly closing levels of the FTSE 100 and FTSE 250, January 2015 to October 2026. The shaded area marks the out-of-sample test period

Table 1 reports summary statistics for weekly returns. Both series are strongly non-normal, with negative skewness and excess kurtosis above 12, driven by the March 2020 crash of about 18.6% in a single week. The Ljung–Box statistic shows no significant autocorrelation in FTSE 100 returns but significant autocorrelation in FTSE 250 returns ($Q(10) = 23.34$, $p < 0.01$). By contrast, squared returns are highly autocorrelated in both indices and the ARCH-LM test rejects homoscedasticity at the 1% level, which confirms volatility clustering and justifies GARCH modelling. The ADF test does not reject a unit root in log prices for the FTSE 100 and rejects it only at the 10% level for the FTSE 250, while returns are clearly stationary; modelling returns rather than price levels is therefore appropriate.

Table 1 Descriptive statistics of weekly log returns, January 2015 to October 2026

Statistic	FTSE 100	FTSE 250
Observations (weeks)	615	615
Mean weekly return (%)	0.081	0.067
Standard deviation (%)	2.073	2.323
Minimum (%)	−18.59	−18.62

Statistic	FTSE 100	FTSE 250
Maximum (%)	7.59	15.16
Skewness	−1.526	−0.935
Excess kurtosis	12.77	12.23
Annualised return (%)	4.19	3.51
Annualised volatility (%)	14.95	16.75
Jarque–Bera	4418.6***	3922.9***
Ljung–Box Q(10), returns	9.74	23.34***
Ljung–Box Q(10), squared returns	103.89***	298.18***
ARCH-LM (4 lags)	84.91***	139.28***
ADF, log price (constant and trend)	−2.45	−3.20
ADF, returns (constant)	−25.29	−13.96

Notes: *** denotes significance at the 1% level. Tests follow Jarque and Bera (1980), Ljung and Box (1978), Engle (1982) and Dickey and Fuller (1979). ADF lag length chosen by AIC. Asymptotic ADF critical values with constant and trend are −3.96 (1%), −3.41 (5%) and −3.13 (10%); with constant only, −3.43, −2.86 and −2.57.

4.2 GARCH estimates

Table 2 reports full-sample estimates of the AR–GARCH and AR–GJR-GARCH models with Student-t errors. Three results stand out. First, volatility is highly persistent: in the GARCH model, the half-life of a volatility shock is about 11 weeks for the FTSE 100 and 8 weeks for the FTSE 250. Second, volatility responds asymmetrically. In the GJR model the coefficient on positive shocks (α) falls to its lower bound of zero, while the asymmetry coefficient (γ) is about 0.25 and significant for both indices, so volatility rises after falls in the market but barely responds to rises of the same size. Third, the low degrees-of-freedom parameter (ν between 4.4 and 7.1) confirms fat-tailed shocks, supporting the Student-t specification.

The GJR model is preferred on every criterion. It has lower AIC and BIC for both indices, and it removes the remaining ARCH effects that the symmetric GARCH leaves in the standardised residuals (for the FTSE 100, ARCH-LM falls from 45.09 to 5.82). Fig. 2 plots the fitted GJR volatility, which spikes at the 2020 pandemic crash and rises again during the 2022 UK mini-budget crisis and the April 2025 tariff shock.

Table 2 Full-sample AR–GARCH and AR–GJR-GARCH estimates with Student-t errors

Parameter	FTSE 100 GARCH	FTSE 100 GJR	FTSE 250 GARCH	FTSE 250 GJR
μ (mean)	0.182*** (0.058)	0.127** (0.059)	0.178*** (0.064)	0.109* (0.066)

Parameter	FTSE 100 GARCH	FTSE 100 GJR	FTSE 250 GARCH	FTSE 250 GJR
ϕ (AR(1))	−0.092** (0.039)	−0.090** (0.040)	−0.057 (0.043)	−0.048 (0.043)
ω	0.239 (0.148)	0.544** (0.229)	0.401** (0.160)	0.307*** (0.115)
α (ARCH)	0.064** (0.032)	0.000 (boundary)	0.163*** (0.046)	0.000 (boundary)
β (GARCH)	0.873*** (0.062)	0.724*** (0.094)	0.750*** (0.062)	0.799*** (0.048)
γ (asymmetry)	–	0.246** (0.104)	–	0.244*** (0.067)
ν (t degrees of freedom)	4.438*** (0.780)	4.905*** (0.909)	5.860*** (1.227)	7.144*** (1.710)
Persistence	0.937	0.848	0.913	0.921
Half-life of a shock (weeks)	10.7	4.2	7.6	8.4
Log-likelihood	−1229.1	−1223.1	−1269.4	−1258.8
AIC	2470.1	2460.1	2550.7	2531.7
BIC	2496.6	2491.1	2577.3	2562.6
Ljung–Box Q(10), standardised residuals	9.87	10.73	10.08	7.52
Ljung–Box Q(10), squared standardised residuals	49.65***	9.05	16.24*	8.80
ARCH-LM (4), standardised residuals	45.09***	5.82	13.19**	6.06

Notes: Standard errors from the inverse Hessian in parentheses. ***, ** and * denote significance at the 1%, 5% and 10% levels. “Boundary” indicates that the estimate lies at the non-negativity constraint, where the standard error is not defined. Persistence is $\alpha + \beta + \gamma/2$.

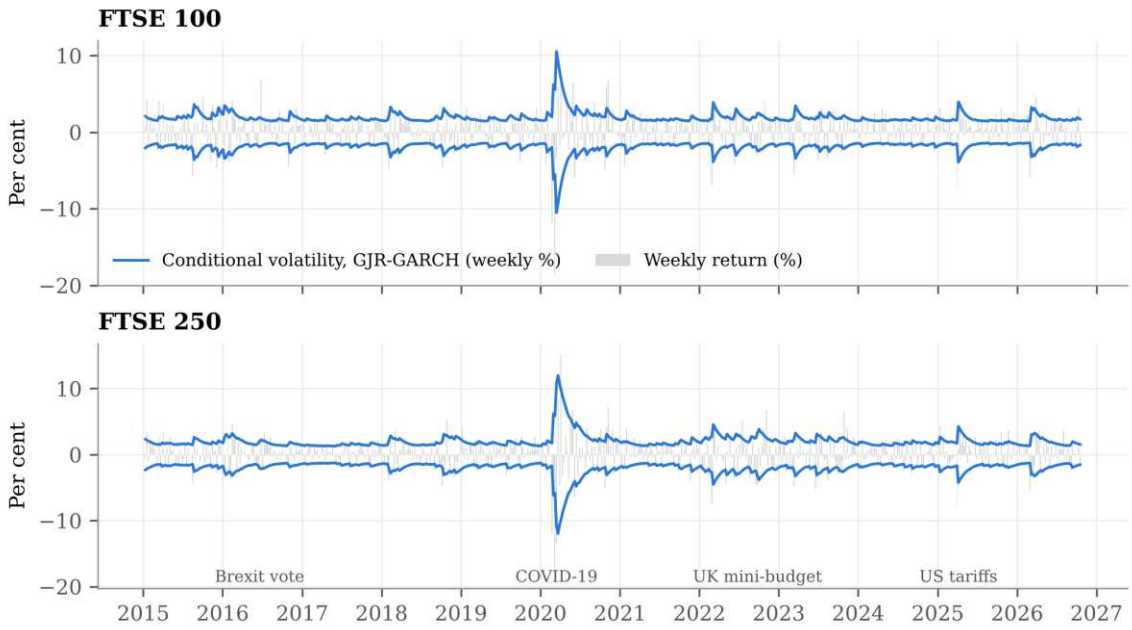


Fig. 2 Weekly returns (grey bars) and GJR-GARCH conditional volatility shown as ± one conditional standard deviation, full sample

4.3 Predicted versus actual prices

Table 3 compares predicted and actual index levels over the 251 test weeks. Every model produces price forecasts that look highly accurate: the mean absolute percentage error is about 1.3% for the FTSE 100 and 1.6% for the FTSE 250, and Fig. 3 shows predicted price paths lying almost on top of the actual paths. This accuracy, however, comes almost entirely from the persistence of prices. The random walk, which simply repeats last week's price, achieves an RMSE of 144.3 points for the FTSE 100, and no model improves on it by more than 0.23 points, or 0.2%. For the FTSE 250, the random walk has the lowest RMSE of all eight models.

Table 3 Accuracy of one-week-ahead price forecasts, January 2022 to October 2026 (251 weeks)

Model	FTSE 100 RMSE (points)	FTSE 100 MAE (points)	FTSE 100 MAPE (%)	FTSE 250 RMSE (points)	FTSE 250 MAE (points)	FTSE 250 MAPE (%)
Random walk	144.3	108.1	1.294	427.4	329.2	1.624
Historical mean	144.1	107.6	1.289	428.2	330.1	1.629
ARIMA	144.5	107.7	1.290	430.2	333.0	1.639
AR–GARCH	144.3	106.7	1.279	431.7	331.0	1.634
AR–GJR-GARCH	144.3	107.1	1.283	430.2	329.8	1.627

Model	FTSE 100 RMSE (points)	FTSE 100 MAE (points)	FTSE 100 MAPE (%)	FTSE 250 RMSE (points)	FTSE 250 MAE (points)	FTSE 250 MAPE (%)
Predictive regression	147.1	110.1	1.318	431.0	331.0	1.631
Random forest	150.8	112.6	1.344	434.2	336.5	1.658
Neural network	144.2	107.4	1.285	428.3	331.6	1.636

Notes: Price forecasts are P(t) × exp(r̂(t+1)/100). RMSE and MAE are in index points.

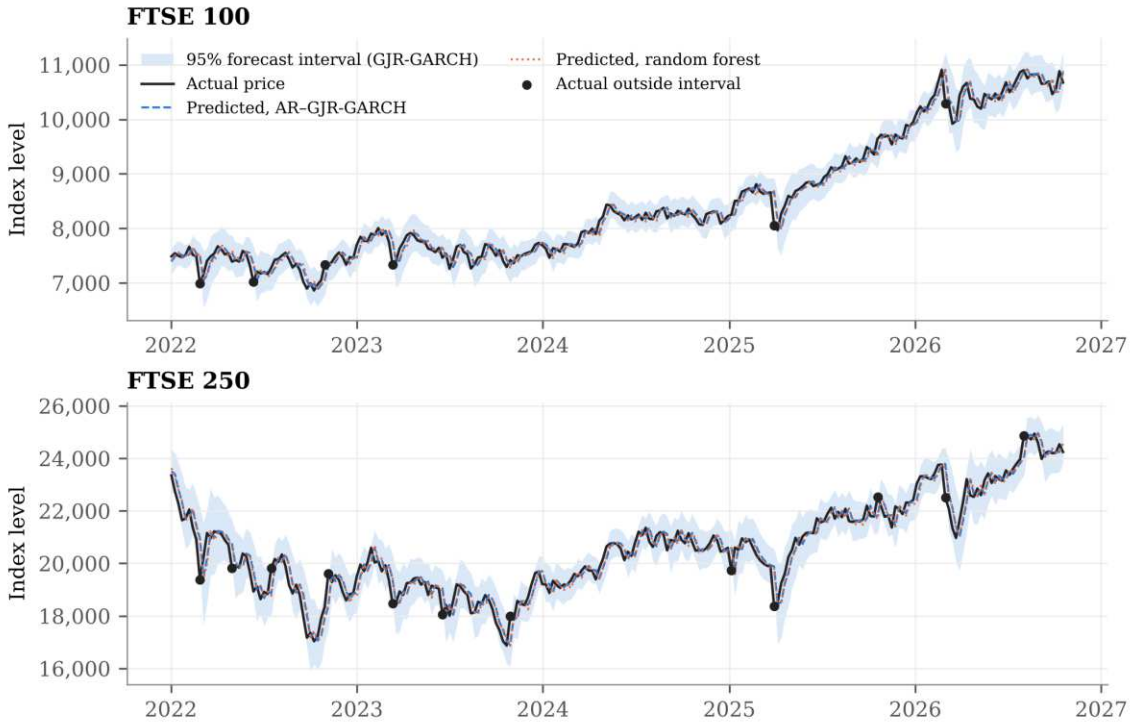


Fig. 3 Actual and predicted weekly index levels in the test period. The shaded band is the GJR-GARCH 95% forecast interval; dots mark weeks in which the actual price fell outside it

Fig. 3 also shows where predictions fail. The actual price leaves the 95% interval in 6 of 251 weeks for the FTSE 100 (2.4%) and 12 weeks for the FTSE 250 (4.8%), close to the 5% expected of a well-calibrated interval. These weeks cluster around abrupt market moves: the February 2022 invasion of Ukraine, the rebound after the autumn 2022 gilt crisis, the March 2023 banking stress and the April 2025 tariff sell-off. In each case the predicted price lags one week behind the actual move: the models follow the market rather than anticipate it. The interval itself, however, widens appropriately after each shock, which shows that the GARCH component captures the changing *size* of risk even when no model captures the *direction* of prices.

4.4 Statistical tests of return forecasts

Because price errors are dominated by the index level, Table 4 evaluates forecasts of weekly returns. Fig. 4 summarises the main result: relative to the random walk, the RMSE ratio lies between 0.998 and 1.044 for every model in both indices, so no model reduces forecast error by more than 0.2%.

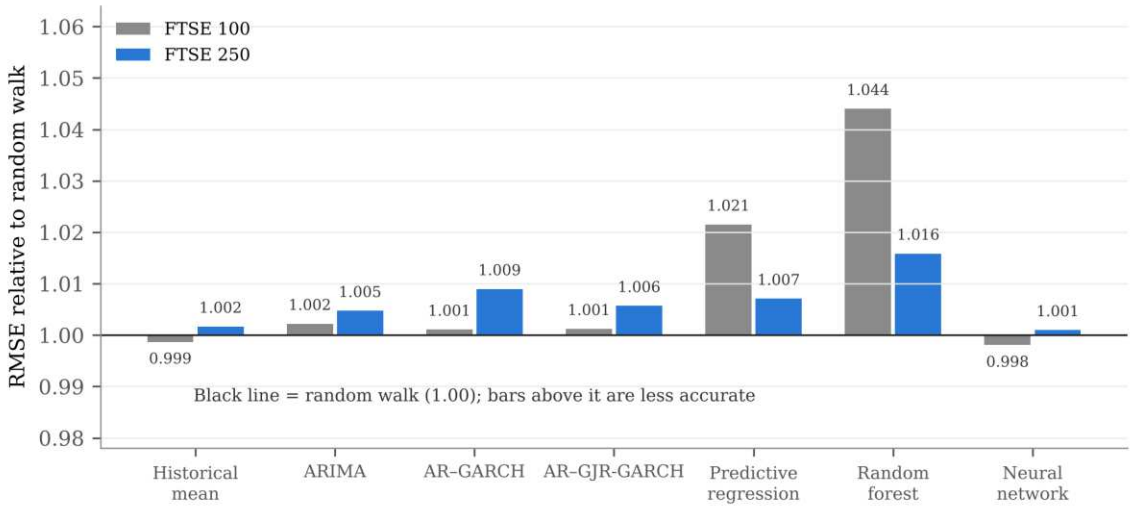


Fig. 4 Root mean squared error of weekly return forecasts relative to the random walk, test period. Values above 1 indicate a less accurate forecast than the random walk

Table 4 Accuracy and predictive-ability tests for weekly return forecasts, test period

Panel A: FTSE 100

Model	RMSE (%)	RMSE ÷ random walk	R^2(OS) vs historical mean (%)	Diebold–Mariano vs random walk	Clark–West (p)	Direction hit rate (%)	Pesaran–Timmermann (p)
Random walk	1.717	1.000	−0.27	–	–	–	–
Historical mean	1.715	0.999	0.00	−0.70	0.171	55.8	–
ARIMA	1.721	1.002	−0.71	0.69	0.648	58.6	0.017
AR–GARCH	1.719	1.001	−0.48	0.12	0.180	56.6	0.128
AR–GJR-GARCH	1.719	1.001	−0.51	0.17	0.252	56.2	0.112
Predictive regression	1.754	1.021	−4.62	2.36**	0.893	52.2	0.235
Random forest	1.793	1.044	−9.28	2.79***	0.863	44.2	0.957
Neural network	1.714	0.998	0.13	−0.28	0.147	52.2	0.600

Panel B: FTSE 250

Model	RMSE (%)	RMSE ÷ random walk	R^2(OS) vs historical mean (%)	Diebold–Mariano vs random walk	Clark–West (p)	Direction hit rate (%)	Pesaran–Timmermann (p)
Random walk	2.128	1.000	0.33	–	–	–	–
Historical mean	2.132	1.002	0.00	1.10	0.816	49.4	–
ARIMA	2.138	1.005	−0.61	0.40	0.290	51.0	0.366
AR–GARCH	2.148	1.009	−1.46	1.29	0.722	50.2	0.272
AR–GJR-GARCH	2.141	1.006	−0.80	1.26	0.798	48.6	0.706
Predictive regression	2.144	1.007	−1.08	0.42	0.102	52.6	0.198
Random forest	2.162	1.016	−2.86	1.32	0.557	49.8	0.521
Neural network	2.131	1.001	0.14	0.15	0.309	46.6	0.908

Notes: The Diebold–Mariano statistic uses squared errors with the Harvey et al. (1997) correction; a positive value means the model is less accurate than the random walk. ** and *** denote significance at the 5% and 1% levels. The Clark–West test is one-sided (H1: model beats the random walk). The direction hit rate is undefined for the random walk, and the Pesaran–Timmermann test is undefined for the historical mean, which always predicts a rise.

The formal tests confirm the visual impression. Not one model beats the random walk at conventional significance levels in either index, on either the Diebold–Mariano or the Clark–West test. Two machine-learning models are significantly *worse*: for the FTSE 100, the random forest (DM = 2.79, $p < 0.01$) and the predictive regression (DM = 2.36, $p < 0.05$) produce larger errors than simply assuming no change. The neural network has the lowest FTSE 100 RMSE and the only positive out-of-sample R^2 in both indices, but the gains of 0.13% and 0.14% are economically negligible and statistically insignificant.

Direction forecasts tell a similar story. The only significant directional result is the ARIMA model for the FTSE 100, which predicts the sign of the weekly change correctly in 58.6% of weeks (Pesaran–Timmermann $p = 0.017$). Even this exceeds the 55.8% that a naive “always up” forecast would have achieved by only 2.8 percentage points. The random forest predicts the direction of the FTSE 100 correctly in only 44.2% of weeks, worse than a coin toss.

Fig. 5 shows how these differences build up over time, following Welch and Goyal (2008). A line that rises means the model is beating the random walk. For the FTSE 100, the random forest and predictive regression lose ground almost continuously, with the largest losses around the 2022 rate shock and the 2025 tariff sell-off. For the FTSE 250, every model sits below zero for most of the test period; the neural network ends closest to zero.

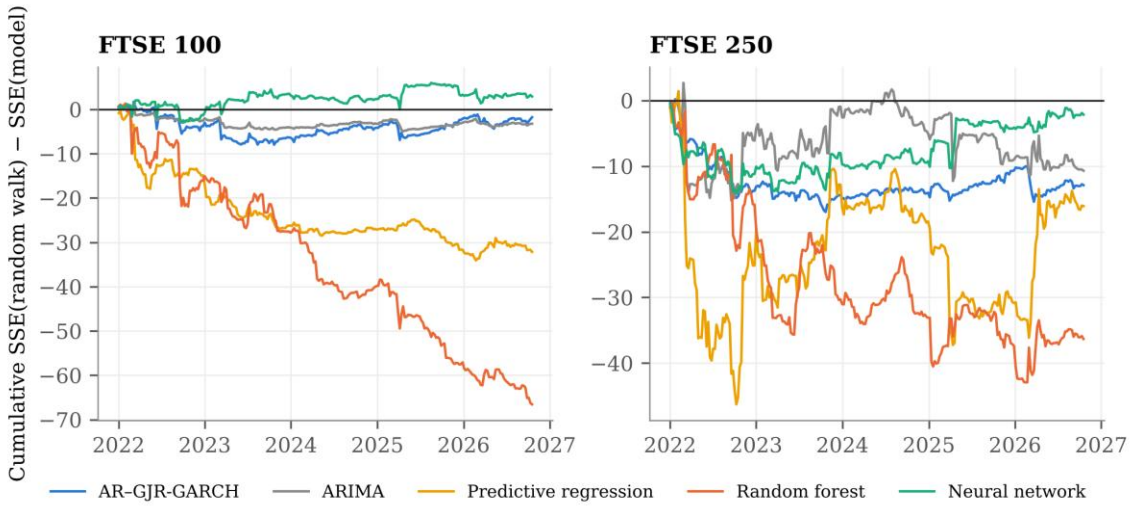


Fig. 5 Cumulative sum of squared forecast errors of the random walk minus those of each model, test period. A rising line indicates that the model is outperforming the random walk

4.5 Forecast bias: Mincer–Zarnowitz regressions

Table 5 regresses actual weekly returns on predicted returns. An unbiased and informative forecast would have an intercept near zero, a slope near one and a meaningful R^2. None of the models meets this standard. The R^2 never exceeds 1.22%, so predicted returns explain almost none of the variation in actual returns. Slopes are mostly near zero or negative, meaning that a higher predicted return is not followed by a higher actual return. Joint unbiasedness is rejected at the 5% level for the predictive regression and random forest on the FTSE 100, and for both GARCH models and the random forest on the FTSE 250. The neural network is the only model whose forecasts are never rejected as biased, but its R^2 of 0.11–0.18% shows that this reflects forecasts that barely move away from zero rather than genuine predictive content.

Table 5 Mincer–Zarnowitz regressions of actual on predicted weekly returns, test period

Model	FTSE 100 a	FTSE 100 b	FTSE 100 Wald (a=0, b=1) p	FTSE 100 R^2 (%)	FTSE 250 a	FTSE 250 b	FTSE 250 Wald p	FTSE 250 R^2 (%)
ARIMA	0.218 (0.107)	−1.929 (1.446)	0.067	0.58	0.001 (0.131)	0.287 (0.404)	0.210	0.18
AR–GARCH	0.145 (0.115)	0.011 (0.443)	0.083	0.00	0.350 (0.217)	−1.854 (1.113)	0.025	0.85
AR–GJR-GARCH	0.147 (0.101)	0.000 (0.462)	0.077	0.00	0.336 (0.186)	−3.017 (1.524)	0.029	1.22
Predictive regression	0.149 (0.092)	−0.501 (0.404)	0.001	0.55	0.002 (0.130)	0.377 (0.316)	0.122	0.82

Model	FTSE 100 a	FTSE 100 b	FTSE 100 Wald (a=0, b=1) p	FTSE 100 R^2 (%)	FTSE 250 a	FTSE 250 b	FTSE 250 Wald p	FTSE 250 R^2 (%)
Random forest	0.148 (0.091)	−0.331 (0.262)	0.000	0.60	0.014 (0.135)	−0.061 (0.429)	0.047	0.01
Neural network	0.114 (0.104)	0.363 (0.539)	0.400	0.11	−0.071 (0.166)	0.638 (0.985)	0.664	0.18

Notes: Regression r(t+1) = a + b r̂(t+1) + u(t+1). Newey–West standard errors in parentheses. The Wald test is of the joint hypothesis a = 0 and b = 1.

4.6 Volatility and risk forecasts

The picture changes completely when the target is the size of price moves rather than their direction. Table 6 shows that the GJR-GARCH model produces the most accurate volatility forecasts for both indices under both loss functions. For the FTSE 250 its advantage over the RiskMetrics EWMA benchmark is statistically significant (DM = −2.93, $p < 0.01$); for the FTSE 100 it is directionally similar but not significant (DM = −1.08). The 12-week rolling variance is the least accurate model and is significantly worse than EWMA for the FTSE 100. GJR forecasts also track the realised range far more closely: their squared correlation with realised volatility is 13.2% for the FTSE 100 and 24.5% for the FTSE 250, compared with 2.0% and 3.4% for EWMA. Fig. 6 shows why: GJR forecasts jump immediately after a large fall and decay within weeks, whereas EWMA reacts more slowly and stays elevated for longer.

Table 6 Accuracy of one-week-ahead volatility forecasts, test period

Volatility model	FTSE 100 MSE	FTSE 100 QLIKE	FTSE 100 DM vs EWMA (QLIKE)	FTSE 250 MSE	FTSE 250 QLIKE	FTSE 250 DM vs EWMA (QLIKE)
GARCH(1,1)-t	10.08	0.440	−0.20	22.61	0.458	−1.72*
GJR-GARCH(1,1)-t	9.78	0.420	−1.08	19.63	0.414	−2.93***
EWMA (λ = 0.94)	9.87	0.445	–	24.46	0.492	–
12-week rolling variance	12.59	0.541	3.03***	28.85	0.529	1.18

Notes: The realised variance proxy is the Parkinson (1980) range estimator. Lower MSE and QLIKE indicate more accurate forecasts. A negative Diebold–Mariano (DM) statistic means the model is more accurate than EWMA. * and *** denote significance at the 10% and 1% levels.

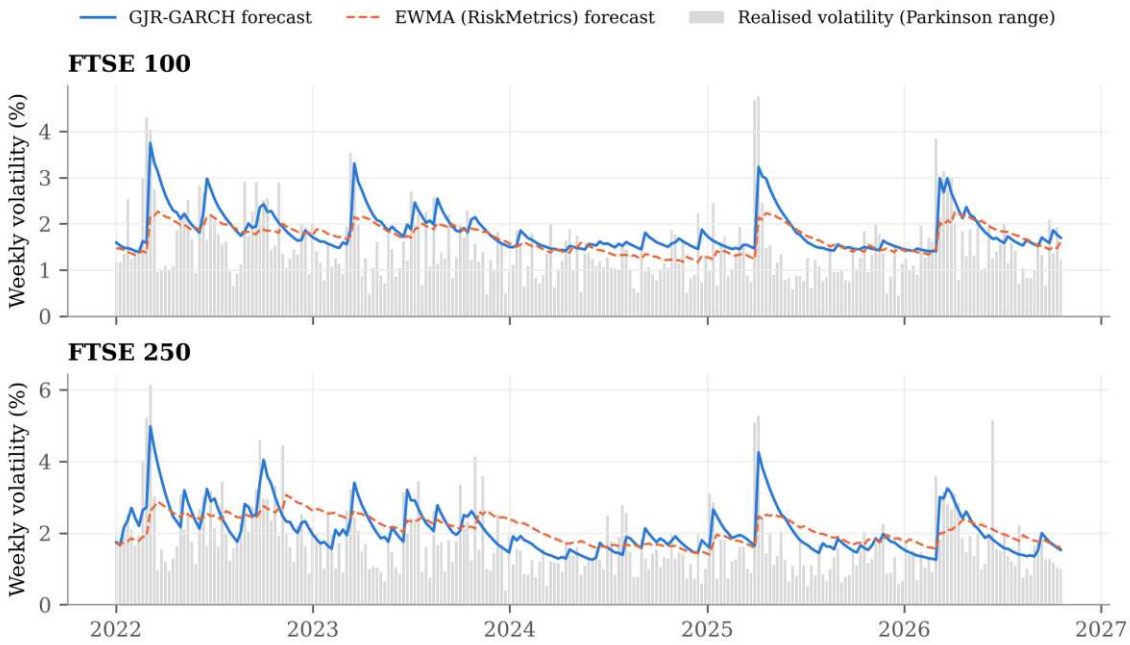


Fig. 6 Realised weekly volatility (Parkinson range estimator, grey bars) and one-week-ahead forecasts from the GJR-GARCH and EWMA models, test period

The Value-at-Risk backtests in Table 7 reach the same conclusion. Both GARCH models produce correctly calibrated 95% and 99% VaR for the FTSE 100, with breach rates of 4.8% and 1.6%. For the FTSE 250 the symmetric GARCH breaches its 95% VaR in 8.0% of weeks and is rejected by the Kupiec test ($p = 0.046$), whereas the GJR model's 7.2% breach rate is not rejected ($p = 0.137$). Allowing volatility to react more strongly to falls therefore matters most for the mid-cap index, which fell further in the 2022 sell-off.

Table 7 Value-at-Risk backtests, test period (251 weeks)

VaR model	FTSE 100 breaches (expected)	FTSE 100 breach rate (%)	FTSE 100 Kupiec p	FTSE 250 breaches (expected)	FTSE 250 breach rate (%)	FTSE 250 Kupiec p
GARCH-t, 95%	12 (12.6)	4.8	0.873	20 (12.6)	8.0	0.046
GJR-GARCH-t, 95%	12 (12.6)	4.8	0.873	18 (12.6)	7.2	0.137
GARCH-t, 99%	4 (2.5)	1.6	0.384	5 (2.5)	2.0	0.164
GJR-GARCH-t, 99%	4 (2.5)	1.6	0.384	5 (2.5)	2.0	0.164

Notes: A breach occurs when the actual weekly return falls below the VaR forecast. The Kupiec (1995) test has the null hypothesis that the breach rate equals the nominal rate.

4.7 Economic value

Table 8 and Fig. 7 test whether forecasts would have helped an investor. For the FTSE 100, which rose strongly after 2024, buy-and-hold earned 7.93% a year with a Sharpe ratio of 0.65, and every forecast-based strategy did worse after costs. The random forest strategy lost 4.20% a year. For the FTSE 250,

which ended the test period close to where it began, buy-and-hold earned only 0.66% a year with a maximum drawdown of 27.8%. Here the ARIMA and predictive-regression strategies earned slightly more (1.97% and 1.36%) with roughly half the drawdown, but they did so by staying out of the market in 44–45% of weeks during the 2022 decline. This is better read as risk reduction in a falling market than as evidence of skill, since the same models showed no significant return predictability in Table 4.

Table 8 Performance of forecast-based trading strategies after transaction costs, test period

Strategy	FTSE 100 annual return (%)	FTSE 100 volatility (%)	FTSE 100 Sharpe	FTSE 100 max drawdown (%)	FTSE 250 annual return (%)	FTSE 250 volatility (%)	FTSE 250 Sharpe	FTSE 250 max drawdown (%)
Buy and hold	7.93	12.30	0.65	−10.57	0.66	15.35	0.04	−27.78
ARIMA	5.89	10.15	0.58	−12.31	1.97	10.93	0.18	−11.22
AR–GARCH	4.27	11.82	0.36	−15.57	−0.20	15.10	−0.01	−31.14
AR–GJR-GARCH	4.31	11.06	0.39	−19.46	−1.93	14.84	−0.13	−31.14
Predictive regression	2.93	8.15	0.36	−11.70	1.36	12.12	0.11	−20.08
Random forest	−4.20	10.09	−0.42	−20.62	−0.55	10.32	−0.05	−20.32
Neural network	3.79	11.19	0.34	−10.56	−1.55	14.36	−0.11	−26.97

Notes: Each strategy holds the index when the predicted weekly return is positive and cash otherwise, paying 0.10% per change of position. Cash earns zero, so Sharpe ratios are annualised return divided by annualised volatility. Prices are price indices and exclude dividends.

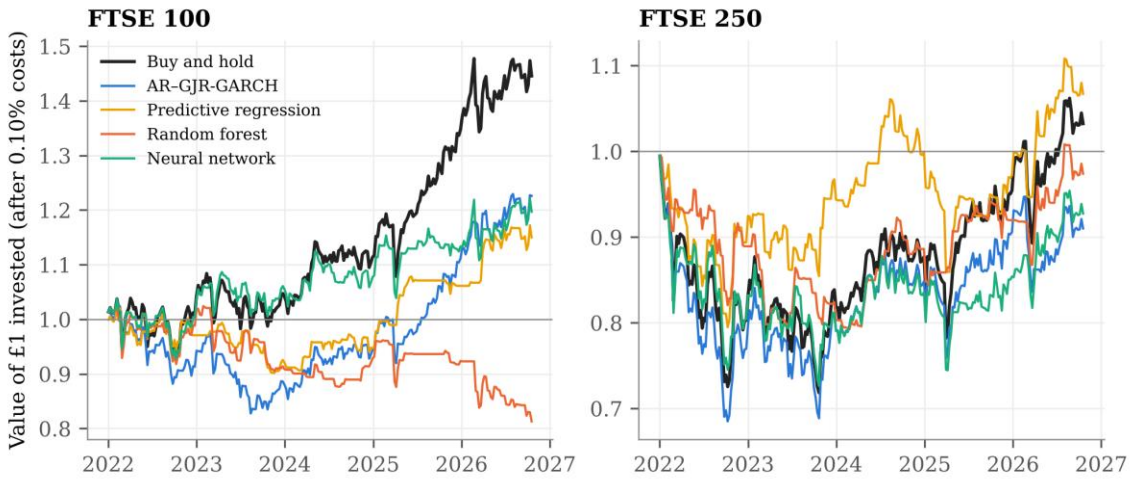


Fig. 7 Value of £1 invested in forecast-based strategies and in buy-and-hold, after 0.10% transaction costs, test period

4.8 Summary of hypothesis tests

Hypothesis	Result	Main evidence
H1: Machine learning beats the random walk for prices	Not supported	No significant improvement for any model; random forest and predictive regression significantly worse for the FTSE 100 (Table 4)
H2: Machine learning beats ARIMA and GARCH for returns	Not supported	Neural network marginally lowest error, but differences are insignificant; random forest least accurate (Tables 3 and 4)
H3: Greater predictability for the FTSE 250	Not supported	In-sample autocorrelation in FTSE 250 returns does not translate into out-of-sample gains; the random walk is the most accurate FTSE 250 model (Tables 1 and 3)
H4: GARCH-family models best for volatility	Supported	GJR-GARCH most accurate for both indices, significantly so for the FTSE 250; only GJR VaR passes all backtests (Tables 6 and 7)
H5: Forecasts add economic value after costs	Not supported	No strategy beats buy-and-hold for the FTSE 100; small FTSE 250 gains reflect time out of a falling market (Table 8)

5 Discussion

5.1 The illusion of accuracy

The central finding is that predicted and actual prices can look almost identical while the forecasts contain no useful information. Every model in this study, including the random forest and neural network, predicts

next week's FTSE 100 level with a mean absolute error of about 1.3%. Yet the same accuracy is achieved by assuming that nothing changes. Because a price index is highly persistent, any model that stays close to last week's price will appear accurate on a price chart. This explains why studies that report price-level errors such as MAPE often present machine-learning forecasts as highly accurate, as noted in reviews of the field (Kumbure et al. 2022; Lin and Lobo Marques 2024). The appropriate test is whether a model predicts *changes* in prices better than the random walk, and on that test none of the eight models succeeds.

5.2 Why machine learning did not help

The results are consistent with weak-form efficiency of the UK index market at weekly frequency (Fama 1970) and with the evidence that in-sample predictability rarely survives out-of-sample evaluation (Welch and Goyal 2008). Three features of the data explain the poor performance of the more flexible models. First, the signal-to-noise ratio is very low: the FTSE 100 shows no significant autocorrelation in returns at all, so there is little for a model to learn. Second, the most flexible model, the random forest, was also the worst, which points to overfitting rather than a lack of model capacity; it learned patterns in the training period that did not recur. Third, the test period contained abrupt regime changes, such as the 2022 rise in interest rates and the 2025 tariff shock, that no model trained on 2015–2021 data could anticipate.

The neural network's near-random-walk behaviour is instructive. With strong regularisation and early stopping, it learned to forecast returns close to zero, which is the most sensible forecast in an efficient market. Its marginal advantage therefore reflects caution, not insight. This result agrees with evidence that machine-learning gains shrink under realistic evaluation (Avramov et al. 2023; Li et al. 2025) and that simple benchmarks are hard to beat (Zeng et al. 2023).

The FTSE 250 result is a useful warning about in-sample evidence. Its returns showed significant autocorrelation over the full sample, which might suggest predictability in a less efficient mid-cap market. Out of sample, however, the random walk was the most accurate model. Predictability that is visible after the event is not necessarily exploitable in real time.

5.3 Where models do add value: risk

In contrast to return forecasting, volatility forecasting works. The GJR-GARCH model forecasts weekly volatility more accurately than historical benchmarks and produces correctly calibrated Value-at-Risk for both indices. This supports the long-standing distinction between the predictability of volatility and the unpredictability of returns (Poon and Granger 2003), and the finding that asymmetric models perform best for equities (Hansen and Lunde 2005). The strong asymmetry estimated here, with volatility responding almost entirely to falls, is consistent with the leverage effect (Black 1976; Glosten et al. 1993). For practitioners, the implication is that the value of quantitative models lies in measuring and managing risk, such as setting position sizes, margins and stop-loss levels, rather than in forecasting the direction of the market.

5.4 Implications for AI adoption in investment

The findings carry three implications for investors and regulators. First, investors should be sceptical of AI tools that present forecast price charts as evidence of accuracy, because a chart of predicted against actual prices cannot distinguish skill from persistence. The minimum standard of evidence is a comparison with the random walk on returns, with formal tests. Second, the economic test matters: the model with

the best statistical record still earned less than buy-and-hold in the rising FTSE 100 market once trading costs were charged. Third, the strong performance of GARCH-type risk models supports the regulatory view that AI and quantitative models are most useful as risk-management aids under human oversight (Bank of England 2025), while the common failure of models during sudden shocks illustrates the risk that many investors relying on similar models could be wrong at the same time.

The results also bear on the emerging literature on large language models. If flexible models using price-based inputs cannot beat the random walk for UK indices, claims that newer AI systems can do so should be tested on data after their training cutoff, against the same benchmark and with the same significance tests (Lopez-Lira et al. 2025; Gao et al. 2026).

6 Conclusion

This study compared predicted and actual prices for the FTSE 100 and FTSE 250 using eight forecasting models, from the random walk to machine-learning methods, over 251 out-of-sample weeks from January 2022 to October 2026. The answer to the question of whether AI predicts UK stock prices better than simple benchmarks is no. Machine-learning forecasts tracked actual prices closely, but no more closely than the random walk, and the random forest and predictive regression were significantly less accurate for the FTSE 100. No model produced unbiased, informative return forecasts, and no forecast-based strategy beat buy-and-hold in the rising FTSE 100 market after costs. By contrast, the asymmetric GJR-GARCH model forecast volatility more accurately than standard benchmarks and produced reliable Value-at-Risk estimates.

The study contributes in three ways. It provides recent UK evidence in a literature dominated by US data. It compares AI and GARCH-family models on price, return, volatility and economic criteria within one walk-forward design. And it shows how an apparent match between predicted and actual prices can mask the absence of predictive skill, which offers a practical test for evaluating claims about AI forecasting tools.

6.1 Limitations

The study has several limitations. It uses two indices rather than individual stocks, where predictability may differ. Weekly frequency smooths intraday and daily patterns that some machine-learning methods exploit. The input set is limited to price, range and volume information, and excludes text, macroeconomic and order-book data. Hyperparameters were fixed in advance rather than tuned, to avoid data snooping, so better-tuned models might perform somewhat differently. The test period of under five years includes unusual episodes such as the 2022 rate shock, and results may differ in calmer markets. Finally, the trading test ignores dividends and interest on cash.

6.2 Future research

Future work could extend the design to FTSE constituents, daily data and longer test periods; add sequence models such as long short-term memory networks and time-series foundation models; test large language models on post-cutoff news; and combine the GJR-GARCH volatility forecasts with machine-learning signals in risk-managed strategies. A further priority is to examine whether investors who rely on AI forecasts of the kind tested here change their trading behaviour, linking forecast performance to market outcomes.

References

Avramov D, Cheng S, Metzker L (2023) Machine learning vs. economic restrictions: evidence from stock return predictability. Manag Sci 69(5):2587–2619. https://doi.org/10.1287/mnsc.2022.4449

Bank of England (2025) Financial Stability in Focus: artificial intelligence in the financial system. Bank of England, London. https://www.bankofengland.co.uk/-/media/boe/files/financial-stability-in-focus/2025/financial-stability-in-focus-artificial-intelligence-in-the-financial-system

Black F (1976) Studies of stock price volatility changes. In: Proceedings of the 1976 Meeting of the Business and Economic Statistics Section. American Statistical Association, Washington, DC, pp 177–181

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econom 31(3):307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Bollerslev T (1987) A conditionally heteroskedastic time series model for speculative prices and rates of return. Rev Econ Stat 69(3):542–547. https://doi.org/10.2307/1925546

Box GEP, Jenkins GM (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Campbell JY, Thompson SB (2008) Predicting excess stock returns out of sample: can anything beat the historical average? Rev Financ Stud 21(4):1509–1531. https://doi.org/10.1093/rfs/hhm055

Cheng Q, Lin P, Zhao Y (2025) Does generative AI facilitate investor trading? Early evidence from ChatGPT outages. J Account Econ 80(2):101821. https://doi.org/10.1016/j.jacceco.2025.101821

Clark TE, West KD (2007) Approximately normal tests for equal predictive accuracy in nested models. J Econom 138(1):291–311. https://doi.org/10.1016/j.jeconom.2006.05.023

Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. J Am Stat Assoc 74(366):427–431. https://doi.org/10.1080/01621459.1979.10482531

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13(3):253–263. https://doi.org/10.1080/07350015.1995.10524599

Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50(4):987–1007. https://doi.org/10.2307/1912773

Fama EF (1970) Efficient capital markets: a review of theory and empirical work. J Finance 25(2):383–417. https://doi.org/10.2307/2325486

Fischer T, Krauss C (2018) Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res 270(2):654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Gao Z, Jiang W, Yan Y (2026) Detecting lookahead bias in LLM forecasts. arXiv preprint arXiv:2512.23847. https://arxiv.org/abs/2512.23847

Glosten LR, Jagannathan R, Runkle DE (1993) On the relation between the expected value and the volatility of the nominal excess return on stocks. J Finance 48(5):1779–1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

Gu S, Kelly B, Xiu D (2020) Empirical asset pricing via machine learning. Rev Financ Stud 33(5):2223–2273. https://doi.org/10.1093/rfs/hhaa009

Hansen PR, Lunde A (2005) A forecast comparison of volatility models: does anything beat a GARCH(1,1)? J Appl Econom 20(7):873–889. https://doi.org/10.1002/jae.800

Harvey D, Leybourne S, Newbold P (1997) Testing the equality of prediction mean squared errors. Int J Forecast 13(2):281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2(5):359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Jarque CM, Bera AK (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ Lett 6(3):255–259. https://doi.org/10.1016/0165-1765(80)90024-5

J.P. Morgan/Reuters (1996) RiskMetrics technical document, 4th edn. J.P. Morgan, New York

Kumbure MM, Lohrmann C, Luukka P, Porras J (2022) Machine learning techniques and data for stock market forecasting: a literature review. Expert Syst Appl 197:116659. https://doi.org/10.1016/j.eswa.2022.116659

Kupiec PH (1995) Techniques for verifying the accuracy of risk measurement models. J Deriv 3(2):73–84. https://doi.org/10.3905/jod.1995.407942

Li WW, Kim H, Cucuringu M, Ma T (2025) Can LLM-based financial investing strategies outperform the market in long run? arXiv preprint arXiv:2505.07078. https://arxiv.org/abs/2505.07078

Lin CY, Lobo Marques JA (2024) Stock market prediction using artificial intelligence: a systematic review of systematic reviews. Soc Sci Humanit Open 9:100864. https://doi.org/10.1016/j.ssaho.2024.100864

Lin CY, Lobo Marques JA, Chan LK (2024) Artificial intelligence and deep learning in stock prediction: a bibliometric review. In: Proceedings of the 20th European Conference on Management Leadership and Governance, pp 297–306. https://doi.org/10.34190/ecmlg.20.1.3003

Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models. Biometrika 65(2):297–303. https://doi.org/10.1093/biomet/65.2.297

Lo AW (2004) The adaptive markets hypothesis: market efficiency from an evolutionary perspective. J Portf Manag 30(5):15–29. https://doi.org/10.3905/jpm.2004.442611

Lopez-Lira A, Tang Y (2023) Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv preprint arXiv:2304.07619. https://arxiv.org/abs/2304.07619

Lopez-Lira A, Tang Y, Zhu M (2025) The memorization problem: can we trust LLMs' economic forecasts? arXiv preprint arXiv:2504.14765. https://arxiv.org/abs/2504.14765

McLean RD, Pontiff J (2016) Does academic research destroy stock return predictability? J Finance 71(1):5–32. https://doi.org/10.1111/jofi.12365

Mincer JA, Zarnowitz V (1969) The evaluation of economic forecasts. In: Mincer JA (ed) Economic forecasts and expectations: analyses of forecasting behavior and performance. National Bureau of Economic Research, New York, pp 3–46

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55(3):703–708. https://doi.org/10.2307/1913610

Parkinson M (1980) The extreme value method for estimating the variance of the rate of return. J Bus 53(1):61–65. https://doi.org/10.1086/296071

Patton AJ (2011) Volatility forecast comparison using imperfect volatility proxies. J Econom 160(1):246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

Pesaran MH, Timmermann A (1992) A simple nonparametric test of predictive performance. J Bus Econ Stat 10(4):461–465. https://doi.org/10.1080/07350015.1992.10509922

Poon SH, Granger CWJ (2003) Forecasting volatility in financial markets: a review. J Econ Lit 41(2):478–539. https://doi.org/10.1257/002205103765762743

Rahimikia E, Ni H, Wang W (2025) Re(Visiting) time series foundation models in finance. arXiv preprint arXiv:2511.18578. https://arxiv.org/abs/2511.18578

Welch I, Goyal A (2008) A comprehensive look at the empirical performance of equity premium prediction. Rev Financ Stud 21(4):1455–1508. https://doi.org/10.1093/rfs/hhm014

Zeng A, Chen M, Zhang L, Xu Q (2023) Are transformers effective for time series forecasting? Proc AAAI Conf Artif Intell 37(9):11121–11128. https://doi.org/10.1609/aaai.v37i9.26317