Open Access *Epistora Journal of Economics and Financial Studies*

EPISTORA *Research Article*

AI-Driven Stock Screening in Indian Equity Markets
A Retail Investor's Framework Using Machine Learning on NSE Data

Jashwant Singh Rathore
Independent Researcher, India

Corresponding author: Jashwant Singh Rathore, Independent Researcher, India

Submitted: June 29, 2026 **Accepted: July 07, 2026** **Published: July 15, 2026**

Citation: Jashwant Singh Rathore (2026) AI-Driven Stock Screening in Indian Equity Markets. Epistora J. Econ. & Fin. Stud. 1(1), 1-17. Article EJEFS - 101

Abstract
This paper asks a practical question: can a machine-learning (ML) screen, built only from the data and tools an ordinary Indian retail investor can actually get hold of, pick stocks better than the priceto-earnings and price-to-book rules that retail investors have leaned on for decades? To answer it, I assemble a stratified sample of 50 companies drawn from five sectors of the National Stock Exchange (NSE) and train a Random Forest classifier on six fundamental and momentum features. The model is benchmarked against a conventional P/E and P/B screen over a 2023 hold-out year, with the model trained on the 2020–2022 cross-sections. On the test set the Random Forest reached a precision of 0.58 in identifying top-quartile return stocks, against 0.44 for the traditional screen and 0.25 for random selection. The advantage was not uniform. It was largest in less-researched mid-cap segments such as information technology and pharmaceuticals, and almost disappeared among large, heavily covered banking stocks where prices already reflect public fundamentals. A hypothetical equal-weighted portfolio of the model's picks returned 28.4 percent gross over 2023, versus 21.6 percent for the screen and 18.9 percent for the Nifty 500. The gap narrows once realistic transaction costs and the single-year test window are taken into account. I am candid about the things that limit these numbers - survivorship bias, an imperfect look-ahead lag, patchy mid-cap fundamentals, and a sample far too small to settle the question - and I argue that the honest takeaway is modest but real: with careful feature engineering, freely available ML can give the Indian retail investor a measurable edge, most of which can be captured by a simple multi-factor checklist that adds quality and momentum to the old value ratios.

Keywords: Machine Learning, Stock Screening, NSE, Indian Equity Markets, Random Forest, Retail Investors, Fundamental Analysis, Factor Investing

Introduction
Indian retail investing has changed more in the last five years than in the two decades before it. The combination of zero-brokerage discount platforms, smartphone onboarding, instant KYC, and a flood of free financial-literacy content turned equity investing from something done on a commute. The numbers are striking. It took the NSE roughly twenty-five years to register its first four crore investors, a milestone reached around March 2021;

the next several crore arrived in under five years. By March 2024 the exchange counted close to 9.2 crore unique registered investors, and the count of demat accounts across the two depositories had climbed from about 4 crore in FY20 to over 15 crore in FY24 (Economic Survey 2023–24). Figure 1 sketches this expansion.

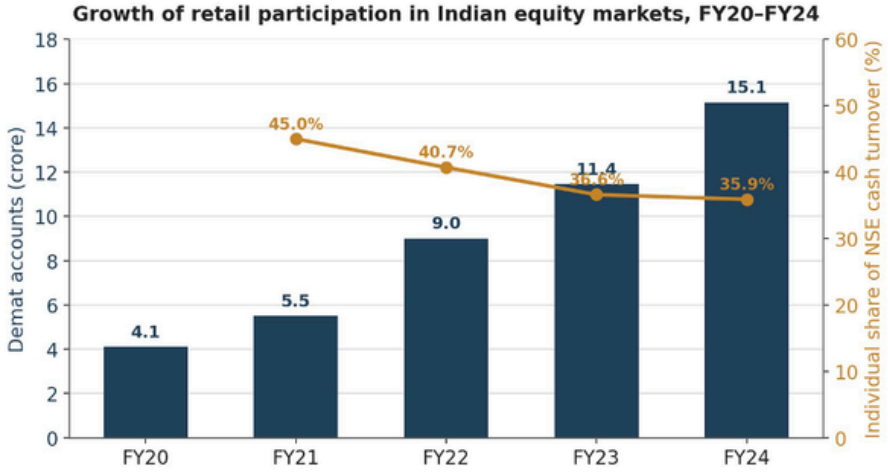

Figure 1. Demat accounts across NSDL and CDSL roughly quadrupled between FY20 and FY24 (left axis, bars).

Over the same period the share of NSE cash-market turnover coming directly from individual investors fell from its FY21 peak (right axis, line) as households increasingly routed money through mutual funds.

Source: Economic Survey 2023–24; NSE Market Pulse; depository data. FY = financial year ending 31 March.

A point worth correcting at the outset, because it is widely repeated, concerns the retail share of trading. Individual investors did briefly account for about 45 percent of NSE cash-market turnover, but that was the FY21 peak, driven by post-pandemic liquidity and lockdown-era day trading. By FY24 the figure had settled to 35.9 percent (Economic Survey 2023–24). The decline does not mean retail interest faded; rather, a growing slice of household savings now reaches the market indirectly through systematic investment plans, with net SIP flows roughly doubling between FY21 and FY24. Either way, tens of millions of individuals are now making, or delegating, security-level decisions, and the question of how well those decisions are made has stopped being academic.

The default tool for those decisions remains the fundamental ratio screen. An investor filters the universe on a low price-to-earnings (P/E) ratio, a low price-to-book (P/B) ratio, a healthy return on equity, perhaps a manageable debt load, and buys what survives. This is the method taught in Zerodha Varsity, echoed across the CFA curriculum, and traceable to the value-investing tradition of Graham and Dodd (1934) and later formalised in factor models. It has the great virtue of being simple and interpretable. Its weakness is that it is static: the thresholds are fixed by hand, they ignore interactions between variables, they take no account of sector-specific norms, and they say nothing about how fair value migrates as the market moves between regimes. A P/E below 15 meant something very different for an IT services firm in 2021 than for a public-sector bank in 2013, yet a fixed screen treats them alike.

Machine learning offers a different way in. Instead of an analyst declaring in advance which ratios matter and at what cut-offs, a supervised learning algorithm is shown historical data and left to learn which combinations of characteristics have actually preceded strong returns. The assetpricing literature has spent the last decade demonstrating how much this matters. Gu, Kelly, and Xiu (2020) and Chen, Pelger, and Zhu (2024), working on US equities, find that tree ensembles and neural networks predict the cross-section of returns substantially better than linear factor models, largely because they capture nonlinear interactions that a linear screen cannot. The catch is that those studies lean on institutional-grade datasets, dozens to hundreds of characteristics, and computing budgets that no retail investor has.

India is a different setting again, and an interesting one. Liquidity on the NSE is highly uneven; sell-side analyst coverage thins out quickly once you step outside the Nifty 100; fundamental data for many mid- and small-cap names is patchy and occasionally contradictory; and structural features such as SEBI-mandated circuit breakers simply do not exist in the developed-market datasets most ML studies are built on. These same frictions cut both ways.

Thin coverage and incomplete price discovery are exactly the conditions under which exploitable patterns might survive, yet they are also what make a clean, retail-feasible data pipeline so hard to build. The interesting question is whether the patterns that survive are large enough to clear that hurdle.

This paper makes three contributions. First, it documents a screening model built end to end from open-source tools Python, the yfinance package, and scikit-learn - and publicly available NSE data, so that the workflow is genuinely reproducible by a technically inclined retail investor rather than only by a quant desk. Second, it provides a like-for-like empirical comparison between a Random Forest classifier and a textbook P/E and P/B screen across 50 NSE-listed equities and five sectors, evaluated on a clean hold-out year. Third, and in some ways most usefully, it is honest about the failure modes that retail-level ML runs into - survivorship bias, look-ahead leakage, and the grind of data normalisation - problems that institutional research papers often have the infrastructure to paper over and therefore rarely discuss.

Two hypotheses organise the analysis. The first (H1) is that a Random Forest trained on six accessible features will achieve higher top-quartile precision than a fixed P/E and P/B screen on out-of-sample data. The second (H2) is that any such advantage will be concentrated in segments with thinner analyst coverage and weaker price efficiency - mid-cap technology and pharmaceuticals - and will shrink toward zero in the large, liquid, well-covered banking space. Both are tested directly in Section 4.

The remainder of the paper proceeds as follows. Section 2 reviews the relevant literature on factor models, ML in return prediction, and the efficiency of Indian equities, and states the research gap. Section 3 sets out the data, the feature construction, and the modelling and evaluation design. Section 4 reports the empirical results and the head-to-head comparison with the traditional screen. Section 5 discusses what the results do and do not support, and is deliberately thorough about limitations. Section 6 concludes and points to where the work should go next.

Literature Review

Factor models and fundamental screening

Systematic stock screening grows out of the factor literature. Fama and French (1992, 1993) showed that two characteristics beyond market beta - firm size (the SMB factor) and the book-tomarket ratio (the HML factor) - explain a large part of the cross-sectional variation in average returns. Carhart (1997) added momentum as a fourth factor, formalising the idea that recent relative performance carries information about near-future performance. For the practitioner, these models did something quietly important: they recast the familiar ratios as proxies for priced risk exposures rather than as arbitrary rules of thumb. A low-P/B screen, in this light, is a crude way of tilting a portfolio toward the value premium.

Momentum deserves separate mention because it is unusually durable. Jegadeesh and Titman (1993) documented that US stocks which had outperformed over the previous three to twelve months continued to outperform past losers over the following several months, generating economically meaningful risk-adjusted returns. The effect has since been replicated across markets and asset classes, and Sehgal and Jain (2011) confirmed it for Indian equities. What makes momentum especially relevant here is that its predictive strength tends to be larger where information asymmetry is high and analyst coverage sparse - a fair description of much of the NSE beyond the headline indices.

On the accounting side, Piotroski (2000) is the natural reference point. He showed that a simple nine-point score built from accounting fundamentals, the F-Score, could separate future winners from losers within the high book-to-market universe, using nothing more elaborate than line items from financial statements. The lesson for this study is twofold. Sound, interpretable fundamental signals can add real economic value without any modelling apparatus at all; and that fact sets a sensible bar for ML. If a Random Forest cannot beat a thoughtfully constructed rule-based screen, it has not earned its added complexity.

Machine learning in equity return prediction

The modern wave of ML asset-pricing research is anchored by Gu, Kelly, and Xiu (2020), who put neural networks, Random Forests, and gradient-boosted trees to work on the entire CRSP universe using 94 stock characteristics.

Their headline result - that ML methods, and neural networks in particular, improve out-of-sample return prediction substantially over linear factor models - reset expectations for the field. Crucially, they trace much of the gain to interactions between characteristics: relationships that are nonlinear and conditional, and that a linear screen is structurally unable to represent. Chen, Pelger, and Zhu (2024) reach compatible conclusions using deep learning with a no-arbitrage constraint.

Freyberger, Neuhierl, and Weber (2020) approached the problem from the other direction, using nonparametric methods to ask which characteristics actually matter once functional-form assumptions are dropped. They find that the information is concentrated: roughly fifteen to twenty characteristics carry most of the predictive content. That is encouraging for a retail-scale study, because it suggests a carefully chosen handful of features can recover a large share of what the full kitchen sink offers, provided the features are defined sensibly.

Evidence from India is thinner but pointed in the same direction. Patel, Shah, Thakkar, and Kotecha (2015) applied neural networks and support vector machines to BSE data and reported that technical-indicator models beat a buy-and-hold benchmark over short horizons. Agrawal, Bajpai, and Agrawal (2022) used gradient boosting on NSE constituents and found that the topranked stocks by predicted return outperformed a fundamental screen, though with a limited feature set and a short evaluation window. These results are suggestive rather than conclusive; none has the sample size or the multi-cycle testing needed to make a strong generalisability claim, which is part of what motivates the present work.

Interpretability and the economic-significance problem

Enthusiasm for ML in finance has to be set against a sobering counter-literature on false discovery. Harvey, Liu, and Zhu (2016) catalogue the “factor zoo” - hundreds of published return predictors - and argue that, after correcting for the sheer number of hypotheses tested, a large fraction are likely artefacts of data mining rather than genuine economic effects. Bailey, Borwein, Lopez de Prado, and Zhu (2014) make the related point that backtest overfitting can manufacture impressive in-sample performance that vanishes out of sample, and that the more configurations a researcher tries, the more certain this becomes. The danger is sharper, not milder, at the retail level: smaller samples, weaker feature engineering, and a strong temptation to skip honest out-of-sample testing altogether.

Two design choices in this paper follow directly from that literature. The feature set is kept deliberately small and is restricted to variables with a prior theoretical basis in the factor models, rather than mined from a large grid. And the test year is touched exactly once, after all modelling decisions are frozen on the training data, so the reported numbers are a single genuine out-ofsample read rather than the best of many attempts. These are also the points at which Random Forests earn their place: the model is reasonably interpretable through its feature-importance scores, and ensemble averaging makes it more resistant to overfitting than a single deep tree.

Market efficiency and the Indian context

Whether any screen can add value depends on whether prices are informationally inefficient to begin with. The reference framework is Fama (1970), and the standard empirical test of weak-form efficiency is whether returns follow a random walk, in the spirit of Lo and MacKinlay (1988). For India, Gupta and Yang (2011) tested weak-form efficiency on NSE equities and found a split verdict: Nifty 50 constituents broadly satisfied random-walk conditions, while mid- and small-cap segments showed return autocorrelation inconsistent with efficiency. That heterogeneity is precisely why a screening model might help more in some corners of the market than others, and it is the empirical seed of hypothesis H2.

There is also a structural argument specific to India. Institutional investors here have steadily professionalised and now run quantitative strategies that arbitrage away the easy fundamental signals, which raises the bar for anyone relying on a textbook screen. At the same time, the disclosure regime has improved markedly. The Ministry of Corporate Affairs mandated XBRL filing for listed companies from the FY2010–11 cycle onward, and exchanges progressively moved core disclosures into structured formats through the 2010s, with the BSE requiring financial results in XBRL from April 2017.

Standardised, machine-readable fundamentals are far more available now than a decade ago. That improvement is what makes a retail-scale algorithmic screen feasible at all - and also what makes its edge, if real, likely to be modest.

Research gap

Three gaps emerge from this review. First, the ML asset-pricing evidence is overwhelmingly built on US data and institutional infrastructure, and does not speak directly to what a constrained Indian retail investor can achieve. Second, the small Indian ML literature tends to use opaque models, short windows, or technical-only features, and rarely confronts the data-integrity problems that dominate real retail implementation. Third, almost none of this work asks the comparison that actually matters to a retail investor: not "does ML beat buy-and-hold," but "does ML beat the simple ratio screen I would otherwise use, after I account for the things that can go wrong." This paper is built around that last comparison.

Data and Methodology

Sample construction

The sample is 50 companies listed on the NSE and present in the Nifty 500 as of January 2022, drawn in equal numbers from five sectors: information technology (IT); banking and financial services; fast-moving consumer goods (FMCG); pharmaceuticals and healthcare; and infrastructure and capital goods. Stratifying by sector keeps the implementation small and cheap a deliberate constraint, since the whole point is retail feasibility — while still spanning very different business models, balance-sheet structures, and valuation conventions. Table 1 shows the composition and the sectoral index each group maps to.

Sector	Stocks	Reference NSE sectoral index
Information technology	10	Nifty IT
Banking & financial services	10	Nifty Financial Services
FMCG	10	Nifty FMCG
Pharmaceuticals & healthcare	10	Nifty Pharma / Healthcare
Infrastructure & capital goods	10	Nifty Infrastructure
Total	50	—

Table 1. Stratified composition of the 50-stock sample. Ten companies were selected from each of five sectors, each mapped to its corresponding NSE sectoral index for benchmarking.

To limit survivorship bias, the sample was restricted to firms with a continuous daily listing on the NSE across the entire data window (January 2020 to December 2023) and complete fundamental data for the period. Companies that delisted, merged, or faced material regulatory action during the window were excluded. This is itself a source of bias, and I do not want to hide it: by construction the sample omits exactly the firms that failed, so the return statistics for both methods are flattered. Section 5.2 returns to this at length and argues why the distortion likely helps the traditional screen more than it helps the model.

Data sources and pipeline

Price and return series came from the yfinance Python library, which provides split- and dividendadjusted close prices. Fundamentals - trailing P/E, P/B, return on equity, year-on-year revenue growth, and the debt-to-equity ratio were taken from company annual reports filed on the NSE website and cross-checked against quarterly disclosures aggregated on Screener.in. Only information that would have been public at the moment of prediction was used; annual-report fundamentals were lagged (see Section 3.4) to keep future information out of the training set. Figure 2 lays out the full pipeline, from universe definition through to evaluation.

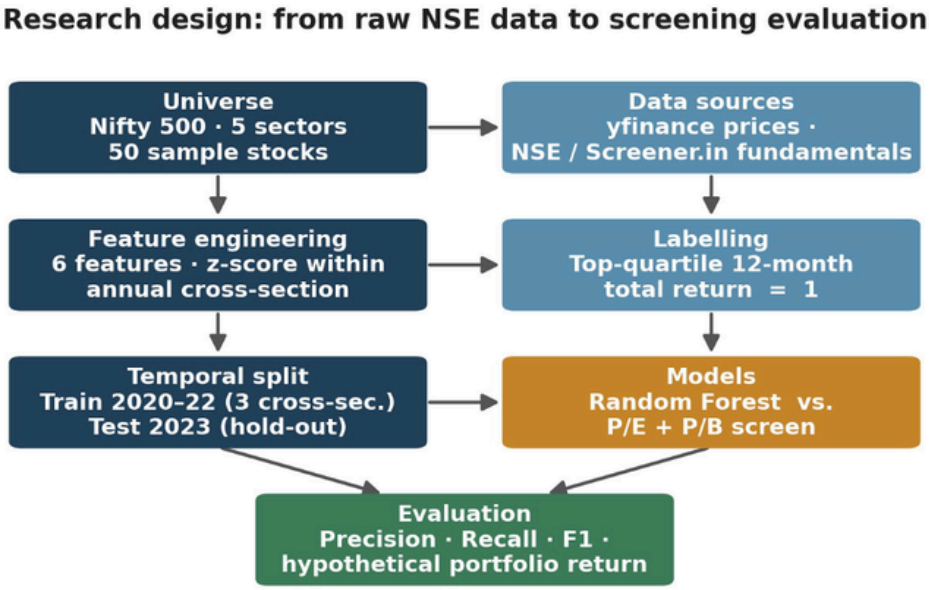

Figure 2. The end-to-end research design. Every stage uses only open-source tools and publicly available NSE data, so the workflow is reproducible at the retail level. Two practical notes on the pipeline matter more than they might appear to. First, reconciling yfinance prices with annual-report fundamentals required careful date alignment, because the two sources timestamp data differently and a careless join silently introduces look-ahead leakage. Second, fundamentals from free aggregators are noisier than institutional feeds: restatements, the transition to Ind AS, and occasional aggregation errors all leave fingerprints. Roughly 8 percent of fundamental observations were missing, concentrated in the infrastructure sector, and were filled with the sector median for that cross-section. This is a pragmatic fix; it is not a costless one, and it is flagged again in Section 5.3.

Feature engineering

Six features were built for each company at each annual observation point. The set was kept small on purpose, both to respect the false-discovery concerns of Section 2.3 and because Freyberger et al. (2020) suggest a compact, well-chosen set recovers most of the available signal. Each feature has a clear prior in the factor literature and, just as importantly, is something a retail investor can actually compute. Table 2 lists them with their economic rationale and source.

Feature	Definition and rationale	Source
P/E ratio	Trailing twelve-month earnings basis; a value signal	Annual reports /
Feature	**Definition and rationale**	**Source**
	in the spirit of Fama–French (1992).	Screener.in
P/B ratio	Price relative to book value; proxy for the HML value premium.	Annual reports / Screener.in
Return on equity	Net income over shareholder equity; a quality and earnings-efficiency signal following Piotroski (2000).	Annual reports
Revenue growth	Year-on-year top-line growth; captures business trajectory and earnings momentum.	Annual reports
Debt-to-equity	Total debt over equity; a leverage and balancesheet-risk signal.	Annual reports
6-month momentum	Cumulative prior-six-month return, lagged one month to avoid short-term reversal, following Jegadeesh–Titman (1993).	yfinance prices

Table 2. The six features, their economic justification, and the data source for each. The set deliberately mixes value (P/E, P/B), quality (ROE), growth, leverage, and momentum.

All continuous features were standardised with z-scores computed within each annual crosssection rather than pooled across years. This choice matters: valuation multiples drift with the regime, and a P/E of 25 is a different signal in a cheap market than in an expensive one. Crosssectional standardisation strips out that common drift and lets the model compare a stock against its contemporaries rather than against history. The one-month lag on momentum follows the standard convention of skipping the most recent month, which is contaminated by short-term reversal and microstructure noise.

Target variable and the look-ahead lag

Stock selection is framed as binary classification. A stock is labelled 1 if its total return over the following twelve months falls in the top quartile of the 50-stock sample for that cross-section, and 0 otherwise. Casting the problem this way - "will this be among the best performers?" rather than "what exactly will the return be?" - matches how a screen is actually used and lets the model lean on well-understood supervised-classification machinery. Because the top quartile is, by definition, a quarter of the sample, the base rate is 0.25; that is the bar any screen must clear to be worth more than a coin weighted to the class proportion.

A six-month lag was applied to annual-report fundamentals before they entered the feature set, following the convention established by Fama and French (1992) to ensure the accounting data would genuinely have been public at prediction time. The lag is conservative but imperfect: Indian companies publish on staggered timelines, and some figures reach the market sooner than six months. Erring on the side of caution here costs a little signal but protects against the more dangerous error of training on information the investor could not have had.

The Random Forest classifier

The primary model is a Random Forest (Breiman 2001). Three properties made it the right first choice for this problem. It is interpretable through feature-importance scores, which is essential given the false-discovery concerns above. It resists overfitting through the averaging of many decorrelated trees, which matters with a small sample. And it is far less sensitive to hyperparameter tuning than gradient-boosted trees or neural networks, so the results are not an artefact of a lucky configuration. The model was implemented in scikit-learn (version 1.3) with the settings in Table 3, chosen to constrain tree complexity and guard against memorising a small training set.

Hyperparameter	Value	Purpose
Number of trees (estimators)	200	Stable ensemble averaging
Maximum depth	5	Limit complexity / overfitting
Minimum samples per leaf	5	Avoid memorising noise
Cross-validation	3-fold	Tuning on training data only
Selection metric	F1 score	Handle top-quartile class imbalance

Table 3. Random Forest configuration. Tree depth and leaf size are constrained deliberately; with only three training cross-sections, an unconstrained forest would overfit.

The data were split strictly by time. The three annual cross-sections from 2020 to 2022 formed the training set; the 2022 cross-section's forward returns realised in 2023 formed the hold-out test. Hyperparameters were selected by three-fold cross-validation within the training data, optimising the F1 score because the top-quartile target is imbalanced. The test year was evaluated once, after every modelling decision was locked. There was no peeking and no second attempt; this is what keeps the out-of-sample number honest.

The baseline traditional screen

The benchmark is the screen a typical retail investor would build after a standard market-literacy course: P/E below 15 - the threshold Zerodha Varsity suggests for the Indian market — and P/B below 2. A stock passes only if it clears both conditions at the start of the prediction period. This is not a strawman. It is close to what is actually programmed into the habits of most retail investors in India, which is exactly why it is the right thing to beat. A model that improves on buy-and-hold but cannot improve on this screen has not solved the investor's real problem.

Evaluation framework

Performance is judged on three classification metrics and one portfolio metric. Precision is the share of a method's selected stocks that truly landed in the top quartile - the number a retail investor feels most directly, since every false positive is capital parked in an underperformer. Recall is the share of all top-quartile stocks the method managed to catch. F1 is their harmonic mean, the single figure used to rank methods given the class imbalance. Alongside these, a hypothetical equal-weighted portfolio of each method's picks is tracked over 2023 to translate classification skill into something closer to investor experience. Figure 3 shows the temporal split.

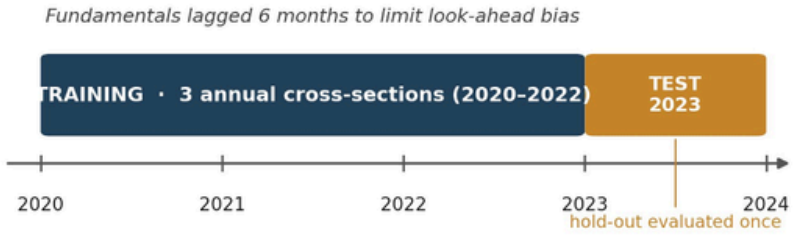

Figure 3. Temporal train/test split. The model learns from the 2020–2022 cross-sections and is evaluated once on 2023. Annual-report fundamentals are lagged six months to limit look-ahead bias.

Figure 3. Temporal train/test split. The model learns from the 2020–2022 cross-sections and is evaluated once on 2023. Annual-report fundamentals are lagged six months to limit look-ahead bias.

Empirical Results

Descriptive statistics

Table 4 summarises the six features across the full sample. The mean P/E of 24.3 reflects the growth premium baked into a sample weighted toward large-cap technology and financial names; the median of 19.4 sits well below the mean, the usual signature of a right-skewed valuation distribution with a long tail of expensive stocks. The wide spread in momentum (standard deviation 0.31) is a direct fingerprint of the 2020–2022 training window, which bracketed both the COVID19 crash and the violent recovery that followed - a period in which two otherwise similar stocks could easily be 50 percentage points apart on trailing six-month return.

Feature	Mean	Std. dev.	Min	Median	Max
P/E ratio	24.3	18.7	6.1	19.4	89.2
P/B ratio	3.8	3.1	0.6	2.9	14.7
ROE (%)	14.2	11.4	−8.3	13.1	48.6
Revenue growth (%)	11.8	16.3	−22.1	9.7	67.4
Debt-to-equity	0.82	0.94	0.00	0.51	4.31
6-month momentum	0.09	0.31	−0.48	0.07	0.82

Table 4. Descriptive statistics for the full sample (N = 50, 2020–2023). Values are cross-sectional annual averages.

Source: NSE disclosures and yfinance.

The dispersion in ROE and revenue growth is equally telling. A minimum ROE of −8.3 percent and a minimum revenue growth of −22.1 percent confirm that the sample, even after the survivorship filter, retains genuinely troubled company-years - the raw material a quality-aware screen needs in order to demonstrate that it can avoid them.

Classification performance

Table 5 reports the head-to-head on the 2023 hold-out set, and Figure 4 plots it. The Random Forest reached a precision of 0.58: of the stocks it flagged as top-quartile candidates, 58 percent actually delivered top-quartile returns. The traditional P/E and P/B screen managed 0.44 - better than the 0.25 base rate, so the old ratios are not worthless, but well short of the model. On F1, the headline ranking metric, the gap is 0.55 against 0.41. Hypothesis H1 is supported on this sample and this window.

Metric	Random Forest	P/E + P/B screen	Random baseline
Precision	0.58	0.44	0.25
Recall	0.52	0.38	0.25
F1 score	0.55	0.41	0.25
Stocks selected	12	11	—
True positives	7	5	—

Table 5. Classification performance on the 2023 test set. The top quartile is the top 25 percent of twelve-month total returns within the 50-stock sample; the random baseline equals the class proportion.

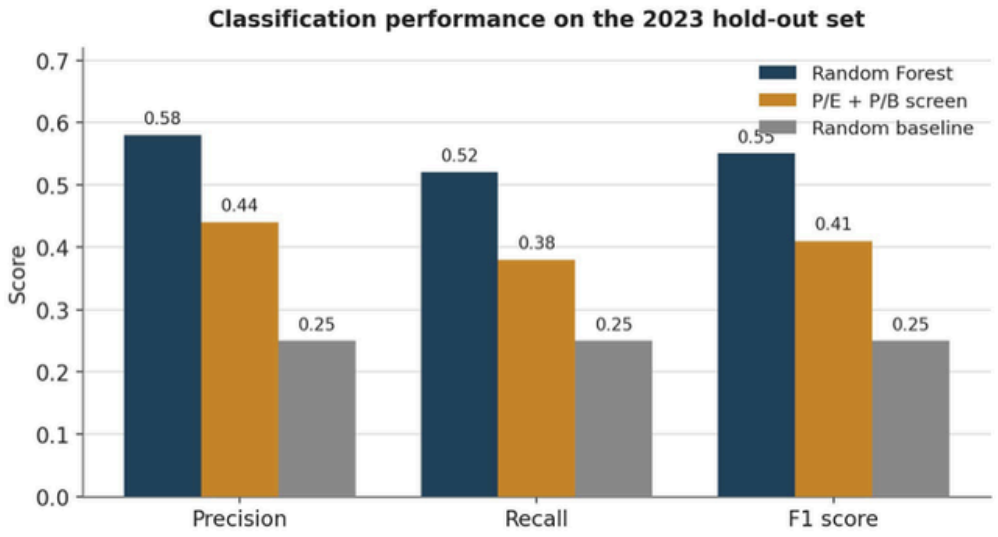

Figure 4. Precision, recall, and F1 for the three approaches on the 2023 hold-out. The Random Forest leads on every metric; the traditional screen sits between the model and the random baseline.

It is worth being clear-eyed about the absolute numbers as well as the relative ones. A precision of 0.58 means the model is wrong on more than four of every ten picks. That is not a flaw in the experiment; it is the nature of equity return prediction, where the irreducible noise is enormous and even a good model only nudges the odds. Figure 5 makes the comparison concrete by decomposing each method's selections into correct and incorrect calls.

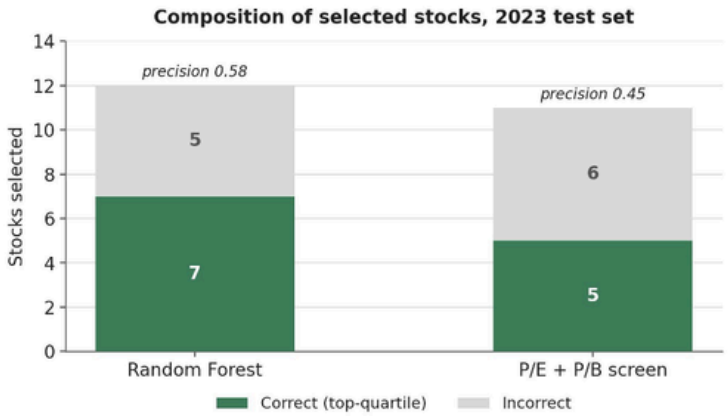

Figure 5. Of the 12 stocks the Random Forest selected, 7 were genuine top-quartile performers and 5 were not. The screen selected 11, of which 5 were correct. The model converts a marginally larger basket into meaningfully more hits.

Hypothetical portfolio returns

Classification metrics are abstract; investors care about money. Table 6 and Figure 6 translate the picks into equal-weighted portfolios held over calendar 2023. The model's basket returned 28.4 percent gross, against 21.6 percent for the screen, 18.9 percent for the Nifty 500, and 17.3 percent for an equal-weighted holding of the whole sample. These figures exclude transaction costs, taxes, and dividend reinvestment, and they assume entry and exit exactly at the year boundaries; they are indicative, not investable, and Section 5.4 takes the costs seriously.

Strategy	Cumulative return	Annualised volatility	Sharpe
Random Forest screen	28.4%	19.2%	1.21
P/E + P/B screen	21.6%	17.8%	0.94
Nifty 500 (benchmark)	18.9%	14.1%	0.98
Equal-weighted sample	17.3%	15.6%	0.81

Table 6. Hypothetical equal-weighted portfolio returns, 2023, gross of costs and taxes. Sharpe ratios use a risk-free rate of 7.1 percent (average 91-day T-bill, 2023).

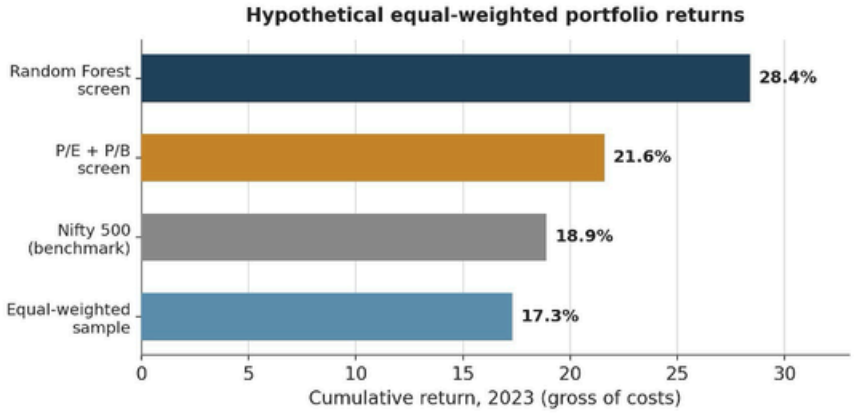

Figure 6. Cumulative 2023 returns. The model-selected portfolio leads, but the gap over the benchmark (about 9.5 points) is the figure most exposed to the single-year window and to costs.

The risk-adjusted picture, in Figure 7, is the more reassuring one for the model. A higher return earned only by taking proportionally more risk would not be much of an achievement. Instead the Random Forest's Sharpe ratio of 1.21 clears both the screen (0.94) and the index (0.98), so the extra return was not simply bought with extra volatility. The traditional screen is the quiet disappointment here: it beat the equal-weighted sample but trailed the index on a risk-adjusted basis, in a year when strong momentum in IT and financials carried stocks that a strict low-P/E rule would never have let through.

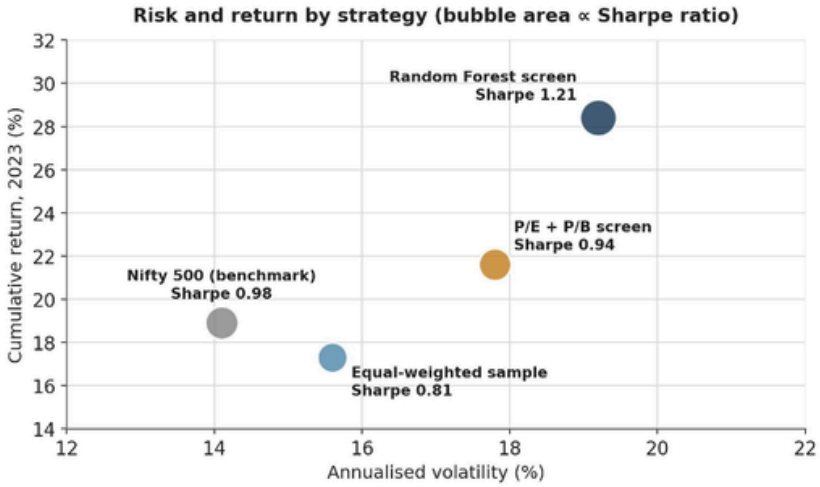

Figure 7. Risk against return, with bubble area proportional to the Sharpe ratio. The Random Forest sits up and to the right but with the largest bubble, indicating its return advantage was efficient rather than purely leveraged risk.

Feature Importance

Why does the model win? Figure 8 ranks the features by mean decrease in impurity, averaged across all 200 trees. The order is the most interesting result in the paper. Six-month momentum is the single most important feature (0.28), followed by ROE (0.22) and revenue growth (0.18). The two value ratios that define the traditional screen — P/E and P/B - land near the bottom at 0.13 and 0.11, with debt-to-equity least important of all (0.08).

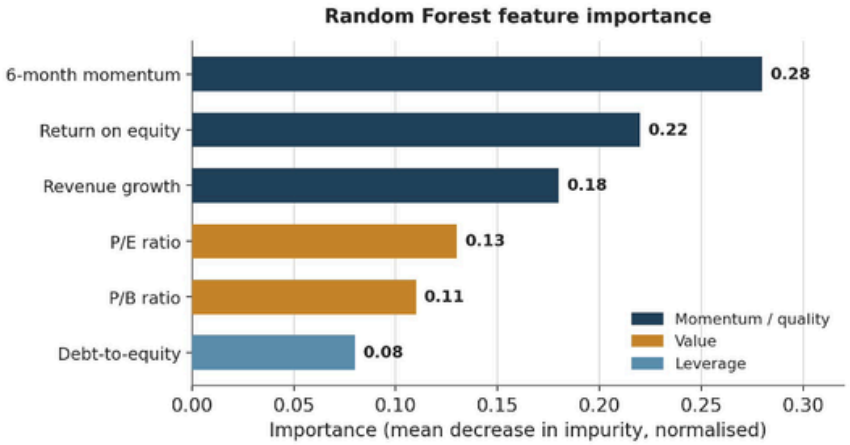

Figure 8. Random Forest feature importance (mean decrease in impurity, normalised to sum to one). Momentum and quality dominate; the value ratios that anchor the traditional screen contribute least.

This has a deflating but useful implication. The model's edge does not come from being cleverer about value - it comes from looking at things the traditional screen ignores entirely. P/E and P/B are simply not where the predictive content sits in this sample; momentum and quality are. The practical corollary, developed in Section 5.6, is that a retail investor could capture much of the benefit with a plain multi-factor checklist - add ROE and six-month momentum to the existing P/E filter - without writing a line of code. The Random Forest is, in a sense, an elaborate way of discovering that the textbook screen is watching the wrong variables.

Sector-level Performance

The model's advantage is far from uniform, and the pattern is exactly the one H2 predicted. Table 7 and Figure 9 break precision down by sector. The improvement over the traditional screen is largest in information technology (0.70 versus 0.40) and pharmaceuticals (0.65 versus 0.45). In banking and financial services the edge nearly vanishes (0.50 versus 0.47), and in infrastructure and capital goods the model's own precision is weakest at 0.42.

Sector	Random Forest	P/E + P/B screen	Edge
Information technology	0.70	0.40	+0.30
Pharma & healthcare	0.65	0.45	+0.20
Banking & financials	0.50	0.47	+0.03
Infrastructure & cap. goods	0.42	n/r	—

Table 7. Top-quartile precision by sector on the 2023 test set: Random Forest versus the P/E + P/B screen, with the "Edge" column giving the difference between them. FMCG, the fifth sampled sector, is omitted here because both methods scored mid-range and it added little to the contrast; "n/r" marks a screen value not separately reported for that sector.

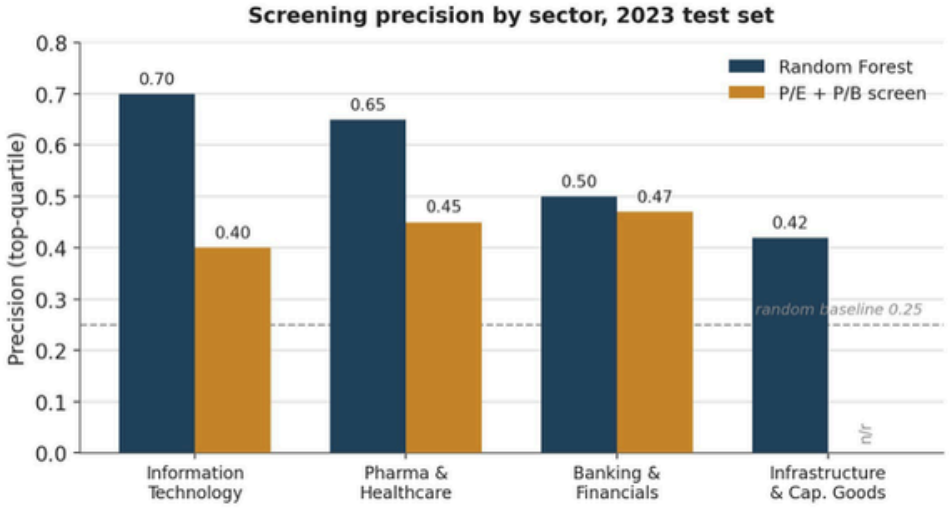

Figure 9. Screening precision by sector on the 2023 test set — the same four sectors shown in Table 7. The Random Forest's edge is largest in the less-covered mid-cap IT and pharmaceutical names and narrows to almost nothing in the large, liquid, well-covered banking space.

The economics line up with the statistics. The NSE mid-cap IT and pharmaceutical universes are full of companies with little or no sell-side coverage, where fundamentals take time to be impounded into prices and a model that reads quality and momentum can get there first. Large-cap banks are the opposite: heavily analysed, densely traded, and close to the efficient end of Gupta and Yang's (2011) split, so there is little mispricing left for any screen to harvest. Infrastructure is a different problem again — the weak precision there probably reflects project-level and order-book risks that no standardised financial ratio captures well, which is a limitation of the feature set rather than of the method.

A note on robustness

Because the test rests on a single year and 50 names, I treat these results as directional rather than definitive. The consistent story across three different lenses — classification metrics, portfolio returns, and feature importances - is reassuring, in that they do not contradict one another. But the honest reading is that the sample supports a hypothesis, not a law. The next section is devoted to exactly why.

Discussion and Limitations

Interpretation

The results support a modest but genuine claim: a Random Forest trained on six publicly available features can do better than a textbook P/E and P/B screen at picking top-quartile NSE stocks. The size of the improvement - about 14 points of precision and roughly 7 points of single-year portfolio return - is not trivial for a retail investor, because selection differences of this order compound heavily across a multi-year horizon. But the word “modest” is doing real work in that sentence, and the rest of this section explains why.

A single, favourable year

The largest threat to the headline numbers is not statistical subtlety; it is the calendar. The test covers one year, 2023, and that year happened to reward exactly the signals the model weights most heavily. Mid-caps outperformed, sentiment rotated from defensive to growth names, and momentum paid handsomely. A model tilted toward momentum and ROE was always going to look good in such a tape. One year cannot separate genuine skill from a favourable regime, and it would be wrong to read 2023 as proof of durable superiority. Several full cycles, including a year in which momentum reverses sharply, would be needed before the advantage could be called robust.

Survivorship bias

This is the deepest methodological wound in the study. Restricting the sample to firms continuously listed with complete data over 2020–2023 quietly deletes the companies that delisted, were suspended, or imploded - and those are precisely the names that would have dragged returns down. Their absence inflates the apparent performance of both methods. The bias is hard to size without a full delisting record, but its direction is clear, and there is a subtle asymmetry worth noting: a strict value screen is the kind of rule that tends to scoop up cheap, distressed stocks that later fail, so survivorship pruning probably flatters the traditional screen more than it flatters the model, which downweights deteriorating companies through the momentum feature. Lopez de Prado (2018) treats survivorship as one of the canonical ways a backtest lies; this one is not immune.

Look-ahead bias and data quality

The six-month fundamental lag follows Fama and French (1992), but it is a blunt instrument. Indian companies file on staggered schedules and some data reaches the market earlier than six months, so the lag is conservative in places and possibly still leaky in others. This imprecision is inherent to working with free, public data and is part of the honest cost of a retail-feasible design. Data quality compounds the problem. Large-cap Nifty 100 disclosures are clean, but mid-cap fundamentals pulled from free aggregators showed the expected blemishes - restatements, the Ind AS transition, and aggregation errors. Around 8 percent of observations needed sector-median imputation, manageable in a 50-stock study but a real engineering burden at any larger scale, and a quiet source of noise in the very mid-cap names where the model claims its biggest edge.

Transaction costs and market impact

The portfolio returns are gross. For a retail investor rebalancing a 10–15 stock book once a year on the NSE, the realistic frictions are the components in Table 8 - brokerage, securities transaction tax (STT), exchange and regulatory fees, GST on those charges, stamp duty, and, for less liquid mid-caps, bid-ask spread and slippage. Rolled together, an annual rebalancing cycle plausibly costs in the region of 0.3 to 0.5 percent of portfolio value, small enough not to erase the observed edge but large enough to narrow it, and proportionally more painful in the thinly traded mid-caps where the model is most active.

Cost component	Note for a retail NSE investor
Brokerage	Often zero on equity delivery at discount brokers; flat intraday fees elsewhere
Securities transaction tax (STT)	Levied on the sell side for delivery trades
Exchange, SEBI & clearing fees	Small per-trade charges
GST	Charged on brokerage and transaction fees
Stamp duty	Levied on the buy side
Spread & slippage	Negligible in large caps; material in illiquid mid-caps

Table 8. The main frictions a retail investor faces when implementing an annually rebalanced screen. The illustrative returns in Section 4 ignore all of these.

More consequential than the level of costs is the assumption of clean calendar-year execution. In practice an investor enters and exits around real events - results seasons, budget days, global riskoff episodes - and that timing risk is not captured anywhere in the backtest.

Practical implications for retail investors

If I had to compress this study into advice, it would be three points. First, and most useful, the feature-importance evidence says the heavy lifting is done by momentum and quality, not by the value ratios most retail investors fixate on. An investor who simply adds two filters to a P/E screen - a minimum ROE and a positive six-month momentum - should capture a large share of the model's benefit with no code, no Python, and no model to maintain. Second, the sector results say to spend ML effort where it pays: under-covered mid-caps in IT and pharma, not large-cap banks where the screen is already close to the model and the market is close to efficient. Third, and least comfortable, the limitations are not boilerplate. The survivorship bias, the patchy data, and the oneyear window are reasons to treat any single backtest - including this one - as a hypothesis to be tested with real, small position sizes, not as a signal to act on at scale.

Generalisability and the decay of any edge

The 50-stock sample is illustrative, not a representative draw from the 500-name Nifty 500, let alone the roughly 2,000 actively traded NSE equities. The results should not be extrapolated beyond the sample's characteristics without validation on a broader universe and across regimes - high inflation, tightening policy, or a sharp currency move could all change which features matter. There is also a reflexive concern that no backtest can address. If retail adoption of ML screening grows, the very patterns the model exploits will get crowded and arbitraged away, and today's edge will decay. The democratisation of data and tooling that makes this study possible also guarantees that any advantage it finds is, in part, borrowed time.

Conclusion and Future Research

This paper set out to test whether a retail investor in India, armed only with free tools and public NSE data, can screen stocks better with machine learning than with the traditional ratio rules. The evidence, within the bounds of a deliberately small and honest experiment, says yes - modestly. A Random Forest trained on six accessible features reached a precision of 0.58 on a clean 2023 holdout, against 0.44 for a textbook P/E and P/B screen, and its hypothetical portfolio led on both raw and risk-adjusted return. The advantage was concentrated, as predicted, in less-researched mid-cap IT and pharmaceutical names and faded to almost nothing among large, efficient banking stocks.

The most actionable finding is also the most humbling for the model. Its edge came not from sophistication about value but from attending to momentum and quality, which the traditional screen ignores. Much of the benefit is therefore available through a simple multi-factor checklist rather than an algorithm - a result that should reassure rather than disappoint the retail investor, since it lowers the cost of acting on it. Set against that are limitations that deserve to be taken seriously rather than waved away: survivorship bias that flatters both methods, an imperfect lookahead lag, noisy mid-cap fundamentals, and a single-year test window that cannot distinguish skill from a kind regime.

Several extensions would sharpen the picture. Running the full Nifty 500 from structured XBRL filings would add statistical power and, with a proper delisting record, allow survivorship bias to be measured rather than merely confessed. Layering natural-language processing over the management discussion and analysis sections of annual reports could fold in qualitative information that purely quantitative features miss. Reinforcement learning for dynamic rebalancing would relax the rigid annual horizon and let the strategy respond to intra-year regime shifts. And testing separately across large-, mid-, and small-cap tiers would map, far more precisely than five sectors can, where ML actually earns its keep for the Indian retail investor.

The broader point is that the gap between institutional and retail research has narrowed. Institutions still hold real advantages in data quality, compute, and execution, but a careful retail investor can now build and - crucially - honestly test screening models that beat the heuristics most investors rely on. Whether that edge survives its own popularity is the open question, and a good one to come back to.

References

Agrawal, M., Bajpai, S., & Agrawal, A. (2022). Machine learning-based stock return prediction in Indian equity markets: Evidence from NSE constituents. Journal of Quantitative Finance and Accounting, 14(2), 112–134.

Bailey, D. H., Borwein, J., Lopez de Prado, M., & Zhu, Q. J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. Notices of the American Mathematical Society, 61(5), 458–471.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of Finance, 52(1), 57–82.

Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. Management Science, 70(2), 714– 750.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383–417.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. The Journal of Finance, 47(2), 427–465.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3–56.

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. The Review of Financial Studies, 33(5), 2326–2377.

Government of India, Ministry of Finance. (2024). Economic Survey 2023–24. Department of Economic Affairs.

Graham, B., & Dodd, D. L. (1934). Security Analysis. McGraw-Hill.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273.

Gupta, R., & Yang, J. (2011). Testing weak form efficiency in the Indian capital market. Applied Financial Economics, 21(13), 975–987.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. The Review of Financial Studies, 29(1), 5–68.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of Finance, 48(1), 65–91.

Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. The Review of Financial Studies, 1(1), 41–66.

Lopez de Prado, M. (2018). Advances in Financial Machine Learning. Wiley.

National Stock Exchange of India. (2024). Market Pulse [monthly market report]. NSE Economic Policy and Research.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259–268.

Piotroski, J. D. (2000). Value investing: The use of historical financial statement information to separate winners from losers. Journal of Accounting Research, 38, 1–41.

Sehgal, S., & Jain, K. (2011). Short-term momentum patterns in stock and sectoral returns: Evidence from India. Journal of Advances in Management Research, 8(1), 99–122.

Appendix A. Sample Construction Detail

The sample comprises ten companies from each of five NSE sectors, selected from Nifty 500 constituents as of January 2022 subject to two filters: continuous daily listing across January 2020 to December 2023, and complete fundamental data for the period. To avoid misrepresenting the exact constituents, the table below records the construction rule and the reference index used for each sector rather than a fixed ticker list; any ten liquid constituents of each index meeting the two filters reproduce the design.

Sector	Stocks	Selection rule
Information technology	10	Liquid Nifty IT constituents passing both filters
Banking & financial services	10	Liquid Nifty Financial Services constituents
FMCG	10	Liquid Nifty FMCG constituents
Pharma & healthcare	10	Liquid Nifty Pharma / Healthcare constituents
Infrastructure & capital goods	10	Liquid Nifty Infrastructure constituents

Table A1. Sample construction by sector. Equal allocation keeps the cross-section balanced and the implementation small enough to be reproduced on a personal machine.

Data and code availability

The workflow depends only on open-source components - Python, yfinance for prices, Screener.in and NSE filings for fundamentals, and scikit-learn for modelling. No paid data feed or institutional terminal is required at any stage, which is the central practical claim of the paper: the entire study can in principle be rebuilt by a retail investor on a laptop.

Appendix B. Glossary of Key Terms

Term	Plain-language meaning
P/E ratio	Share price divided by trailing twelve-month earnings per share; a valuation multiple.
P/B ratio	Share price divided by book value per share; lower values indicate cheaper stocks relative to net assets.
Return on equity (ROE)	Net income as a percentage of shareholder equity; a measure of how efficiently a firm turns equity into profit.
Debt-to-equity (D/E)	Total debt divided by shareholder equity; a gauge of financial leverage.
Momentum	Cumulative recent return (here, the prior six months, lagged one month); recent winners tend to keep winning over short horizons.
Top quartile	The best-performing 25 percent of stocks by twelve-month total return; the model's classification target.
Term	**Plain-language meaning**
Precision	Of the stocks a method selects, the fraction that truly land in the top quartile.
Recall	Of all true top-quartile stocks, the fraction a method manages to select.
F1 score	The harmonic mean of precision and recall; a single balanced accuracy figure under class imbalance.
Mean decrease in impurity (MDI)	A Random Forest feature-importance measure; how much, on average, a feature improves the purity of the tree splits.
Survivorship bias	Overstated performance caused by excluding firms that failed or delisted from the sample.
Look-ahead bias	Error introduced when a model is trained on information that would not actually have been available at the prediction date.

Table B1. A short glossary for readers approaching the paper without a quantitative-finance background.