Forecasting Returns in Thin Markets: A Machine Learning Approach to the Zagreb Stock Exchange

Mislav Šagovac Luka Šikić Petra Palić

2026-06-18

Table of contents

Abstract

Purpose: This paper investigates the out-of-sample predictability of weekly stock returns on the Zagreb Stock Exchange (ZSE), a thin frontier market characterized by low liquidity and concentrated ownership. The study aims to determine whether machine learning algorithms can extract predictive signals in this environment and how liquidity constraints influence forecasting performance.

Methodology: We construct a comprehensive set of over 1,100 predictors from daily OHLCV data spanning from 2000 to 2024, including technical indicators, time-series features, and wavelet decompositions. Four predictive models (Elastic Net, Random Forest, XGBoost, and shallow neural networks) are evaluated using a rigorous nested rolling-window cross-validation framework to prevent information leakage. We assess performance through both statistical metrics and a realistic portfolio backtest.

Results: The results indicate modest directional accuracy (46 to 53 percent), with nonlinear ensemble methods and neural networks consistently outperforming the linear benchmark. Forecast combinations further enhance predictive power. A critical finding is a strong monotonic liquidity gradient: portfolio Sharpe ratios increase from 0.17 for the most liquid stocks to 1.58 for the full stock universe (averaged across models; up to 1.97 for the best individual model), suggesting significantly higher predictability in illiquid securities.

Conclusion: Machine learning models can generate economically significant signals in frontier markets, but this predictability is heavily concentrated in thinly traded stocks. While the apparent risk-adjusted returns are high, practical implementation is likely constrained by transaction costs and market depth, highlighting liquidity as a critical factor in return predictability.

Keywords: machine learning, return predictability, frontier markets, liquidity gradient, ZSE, portfolio backtesting, tree ensembles, neural networks

JEL Classification: C53, G12, G17, C55, G11

---

1 Introduction

The predictability of equity returns remains one of the most consequential questions in empirical finance. While the efficient market hypothesis implies that excess returns should be unforecastable, a growing body of evidence demonstrates that machine learning methods can extract predictive signals from high-dimensional feature spaces that

elude traditional econometric approaches (Gu, Kelly, and Xiu 2020; Bianchi, Büchner, and Tamoni 2021). This predictive capacity has been documented primarily in deep, liquid markets, most notably U.S. equities, where the abundance of data and market efficiency pose a formidable challenge to forecasting models (Kumbure et al. 2022). Whether similar predictability exists in smaller, less liquid markets characterized by thin trading and concentrated ownership structures remains an open empirical question with important implications for asset pricing theory and portfolio management. The potential for machine learning to uncover return predictability in frontier markets is particularly intriguing given theoretical predictions that information asymmetries and trading frictions may generate exploitable patterns that have been arbitraged away in more developed markets.

This paper investigates the out-of-sample predictability of weekly stock returns on the Zagreb Stock Exchange (ZSE), a frontier market that has received limited attention in the machine learning literature despite a substantial body of traditional econometric research (Šego and Škrinjarić 2019). The Croatian equity market presents a compelling laboratory for several reasons. First, its relatively low analyst coverage and institutional participation may permit persistent mispricings that machine learning algorithms can exploit. Second, the prevalence of illiquid securities with infrequent trading creates a natural experiment for examining how predictability varies across the liquidity spectrum, a dimension largely absent from studies focused on developed markets. Third, the ZSE's concentrated sector composition (tourism, energy, banking) and susceptibility to idiosyncratic macroeconomic shocks distinguish it from the diversified market structures typically studied in the literature. Fourth, Croatia's EU accession in 2013 and eurozone entry in 2023 provide a dynamic institutional backdrop that may influence market efficiency over our sample period.

Our empirical framework makes three methodological contributions. First, we construct a comprehensive predictor set comprising over 1,100 features derived from daily OHLCV data, encompassing technical indicators, time-series characteristics (Lubba et al. 2019), explosive behavior tests (Phillips, Shi, and Yu 2015), wavelet decompositions, and univariate forecasting model outputs. This approach connects to the burgeoning literature on the factor zoo and the challenge of selecting relevant predictors from high-dimensional candidate sets (Feng, Giglio, and Xiu 2020; Freyberger, Neuhierl, and Weber 2020). Second, we implement a rigorous nested rolling-window cross-validation scheme with explicit temporal gaps to prevent information leakage, a critical but often neglected consideration in financial machine learning applications (Gu, Kelly, and Xiu 2020). Third, we employ information-theoretic feature selection methods, specifically Joint Mutual Information (JMI) and ReliefF (Kononenko 1994), to identify predictors with maximal incremental information content while controlling for redundancy.

We evaluate predictive performance through two complementary lenses. The first employs standard statistical metrics for both classification (directional accuracy, precision, recall) and regression tasks (MSE, MAE). The second implements a realistic trading simulation in which weekly portfolio positions are constructed based on model forecasts, enabling direct assessment of economic significance through risk-adjusted performance measures. This dual evaluation addresses the well-documented gap between statistical and economic significance in return prediction (Leitch and Tanner 1991).

To structure this investigation, we formulate three testable hypotheses. The first concerns the existence of predictability: H1 posits that machine learning models extract out-of-sample predictive signal from the ZSE feature space, generating directional accuracy and portfolio returns that exceed both random chance and a naive no-change (random-walk) benchmark. The corresponding null hypothesis is that weekly ZSE returns are unpredictable out of sample, consistent with the weak-form efficient market hypothesis. The second concerns the source of predictability: H2 posits that nonlinear models (Random Forest, XGBoost, and the neural network) achieve superior out-of-sample performance relative to the penalized-linear Elastic Net benchmark, implying that predictability arises substantially from nonlinear interactions among predictors rather than from linear factor structure alone. The third concerns the role of liquidity: H3 posits that risk-adjusted predictability increases monotonically as the investable universe is widened to include less-liquid securities, reflecting slower information diffusion in thinly traded stocks. The alternative is that any apparent liquidity gradient is an artifact of unmodelled transaction costs and bid-ask bounce rather than genuine predictability. These hypotheses are evaluated jointly through the statistical and economic criteria described in Section 3.

Section 4 evaluates these hypotheses. In brief, we find qualified support for H1 and H2 — machine learning models, and nonlinear models in particular, generate modest but economically meaningful out-of-sample predictability — and strong descriptive support for the liquidity gradient in H3. We defer all quantitative magnitudes, model comparisons, and robustness considerations to the results section, where they can be interpreted alongside their limitations.

The remainder of this paper proceeds as follows. Section 2 provides a review of the relevant literature on equity return predictability and the application of machine learning methods in financial markets, with particular emphasis on emerging and frontier market contexts. Section 3 details the machine learning pipeline, including data preprocessing, feature engineering, model specification, and the nested cross-validation architecture. Section 4 presents results from both statistical evaluation metrics and portfolio backtests, with particular attention to heterogeneity across liquidity strata. Section 5 concludes with implications for practitioners and directions for future research.

2 Literature Review

The scholarly investigation into the predictability of equity returns has undergone significant evolution, moving from the restrictive principles of the Efficient Market Hypothesis toward modern approaches grounded in machine learning. Although the Efficient Market Hypothesis, formalized by Fama (1970), claims that the consistent generation of abnormal returns is impossible, a large body of empirical evidence demonstrates the existence of persistent market anomalies and systematic risk factors (Fama and French 1993; Carhart 1997). These findings imply that returns are predictable, yet difficult to forecast through simple linear econometric models (Pesaran and Timmermann 2000). The increasing use of machine learning methods emerged precisely because classical models struggle to capture the nonlinear and interactive relationships that characterize high-dimensional financial data (Gu et al. 2020; Bianchi et al. 2021). Machine learning models are well suited to processing this complexity and have proven effective in identifying subtle predictive signals that traditional factor models are unable to detect.

Within this research domain, several model classes have become particularly influential. Tree-based ensemble algorithms, including Random Forests (Breiman 2001) and gradient boosting techniques such as XGBoost (Chen and Guestrin 2016), are valued for their robustness, their capacity to model interactions, and their ability to manage the bias-variance trade-off in noisy environments. Seminal studies show that these models are among the most effective tools for predicting cross-sectional stock returns (Gu et al. 2020). Neural networks also contribute meaningfully to this literature. Even shallow architectures demonstrate strong nonlinear approximation capabilities (Zhang 2000). More advanced deep learning frameworks, including Long Short-Term Memory networks and other recurrent architectures, have shown considerable promise in handling sequential data, learning hierarchical representations, and reducing reliance on manually engineered financial factors (De Prado 2018; Fischer and Krauss 2018). Recent work extends this frontier by adapting Transformer-based architectures to multi-asset forecasting tasks due to their capacity to learn long-range dependencies in financial time series (Liu et al. 2023).

Penalized linear models, particularly Elastic Net (Friedman et al. 2010), offer an important benchmark for the possibility that improvements in predictive accuracy may stem from enhanced variable selection rather than genuine nonlinear structure. This consideration is closely related to the broader challenge of the factor zoo, where a large number of potential predictors necessitates careful dimensionality reduction and control of model complexity (Cochrane 2011). Empirical evidence confirms that high-dimensional predictor sets improve performance only when paired with effective feature selection procedures (Feng et al. 2020). Consequently, information-theoretic approaches such as ReliefF (Kononenko

1994) and Joint Mutual Information (Peng et al. 2005) have become central to filtering redundant or noisy predictors and isolating those with substantial incremental information content.

Extending this body of work, recent research has highlighted the superior performance of machine learning in international and emerging market contexts, where market inefficiencies may amplify predictive signals. For instance, neural networks trained on market-specific data outperform global models, particularly when incorporating U.S.-derived characteristics to enhance predictability in international stock returns (Choi et al. 2025). Similarly, machine learning models reveal that return predictability is strongest in small, illiquid, and less integrated markets, often characteristic of emerging and frontier economies, driven by factors like momentum, earnings yield, and long-term reversal, though translating these into profitable strategies remains challenging due to diminishing capital mobility constraints (Cakici and Zaremba 2024). In emerging markets specifically, tree-based ensembles and neural networks consistently outperform linear benchmarks in predicting cross-sectional returns, generating significant alphas that persist even after accounting for transaction costs and short-selling constraints, suggesting underreaction rather than risk-based explanations (Hanauer and Kalsbach 2023). This is echoed in global studies of stock market anomalies, where feed-forward neural networks and composite predictors exploit nonlinear relationships to achieve high out-of-sample returns, remaining robust to costs up to 300 basis points and challenging efficient market theories across developed, emerging, and frontier markets (Azevedo et al. 2023).

Machine learning applications in frontier and emerging markets also underscore the importance of integrating domain-specific features and addressing practical frictions. For example, combining data envelopment analysis efficiency scores with automatic feature engineering in gradient boosted trees significantly improves prediction accuracy in low-liquidity settings like the Vietnamese stock market (Thanh Nhon et al. 2025). Furthermore, frameworks that embed machine learning directly into portfolio optimization with transaction costs produce an implementable efficient frontier, prioritizing persistent signals like value and momentum while mitigating the pitfalls of high-turnover strategies in illiquid environments (Jensen et al. 2023). Locally, studies on the Zagreb Stock Exchange have applied random forests to predict market trends, demonstrating the feasibility of machine learning for directional forecasting in thinly traded frontier markets, though with a focus on indices rather than individual stocks (Manojlović and Štajduhar 2016).

A growing strand of this literature evaluates deep learning on tabular financial data, where the evidence is notably mixed. While Transformer architectures excel at capturing long-range dependencies in sequential data (Liu et al. 2023), recent systematic compar-

isons find that on heterogeneous tabular feature sets — the setting of cross-sectional return prediction — gradient-boosted trees frequently match or outperform deep networks, which are more sensitive to uninformative features and require larger samples (Grinsztajn, Oyallon, and Varoquaux 2022). This tension motivates our inclusion of both tree ensembles and a neural network within a common evaluation framework. A critical reading of the most closely related prior work further sharpens our contribution. Manojlovic and Stajduhar (2016), the only prior machine-learning study of the ZSE, report encouraging but limited results because they forecast aggregate index direction with a single random-forest model; we extend this to individual securities, a broader model suite, and an explicit liquidity decomposition. Hanauer and Kalsbach (2023) document robust emerging-market predictability that survives transaction costs, yet caution that profitability concentrates in hard-to-trade stocks — a pattern our liquidity-gradient analysis confirms and makes central rather than incidental. The mixed conclusions across these studies stem largely from differences in market depth, the granularity of the prediction target, and whether implementation frictions are modelled; our design addresses these by combining a thin-market setting, security-level weekly forecasts, a nested leakage-controlled validation scheme, and an explicit transaction-cost discussion.

Although most machine learning evidence comes from deep and liquid markets such as the United States (Kumbure et al. 2022), extending this analysis to frontier and emerging markets remains essential. These markets, including the Zagreb Stock Exchange, are characterized by low analyst coverage, concentrated ownership structures, and high information asymmetry, all of which theoretically contribute to slower information diffusion and potentially greater return predictability (Amihud and Mendelson 1986). Previous econometric studies on the Croatian market document the presence of traditional anomalies (Šego and Škrinjarić 2019). Machine learning methods allow for a more detailed examination of the liquidity gradient and the extent to which frictions in thinly traded stocks generate predictable returns. The broader literature also highlights the need to connect statistical predictability with economic relevance (Leitch and Tanner 1991). This requires realistic backtesting procedures that account for market microstructure effects and the possibility of data mining (Sullivan et al. 1999). These considerations are particularly important in frontier markets where transaction costs and limited depth may materially affect realized performance.

3 Empirical Framework

Implementing machine learning for financial time series prediction requires a systematic pipeline that addresses the unique challenges of noisy, non-stationary data with low

signal-to-noise ratios (Gu, Kelly, and Xiu 2020). Our empirical framework comprises five interconnected stages: (i) data acquisition and cleaning, (ii) feature engineering and predictor construction, (iii) model specification, (iv) nested cross-validation with hyperparameter optimization, and (v) portfolio-based performance evaluation.

3.1 Data

We obtain daily trading data for all securities listed on the Zagreb Stock Exchange (ZSE) spanning January 4, 2000 through January 4, 2024. Figure 1 illustrates the time-varying composition of the ZSE universe over the sample period. The number of listed securities expanded from approximately 40 in 2000 to a peak of nearly 100 prior to the 2008 financial crisis, subsequently contracting to approximately 30 active listings by 2019, a level that has remained relatively stable. This contraction reflects broader delisting trends, consolidation in the Croatian corporate sector, and reduced IPO activity following EU accession. Given the limited cross-sectional breadth of the market, we include all securities with sufficient data quality rather than imposing market capitalization or liquidity screens that would further reduce sample size. Throughout this paper we use the terms "securities," "stocks," and "shares" interchangeably to denote ordinary common-equity listings on the ZSE regulated market; the sample comprises exchange-listed equities and excludes bonds, investment funds, and other non-equity instruments.

> **Note**
>
> *Figure placeholder — add* `fig3_cropped.png` *to the* `paper/` *folder and re-render to display the chart.*

Figure 1. *Number of common-equity securities listed on the Zagreb Stock Exchange by year, 2000–2024. Horizontal axis: calendar year. Vertical axis: count of listed securities in the cleaned sample.*

3.2 Data Preprocessing

Raw OHLCV (Open, High, Low, Close, Volume) data were obtained via automated web scraping from official ZSE sources. We implement a multi-stage data cleaning protocol to address common data quality issues in frontier market datasets (Consoli, Recupero, and Saisana 2021). The initial filtering stage addresses five categories of data quality concerns. First, we exclude observations with closing prices below €0.00000001, which typically indicate data errors or placeholder values. Second, we require securities to have at least two years (104 weeks) of trading history to permit reliable feature extraction from rolling windows. Third, we exclude securities for which the ratio of actual trading days to potential trading days falls below 70 percent, as sporadic trading introduces stale-price biases. Fourth, for trading days with missing closing prices, we impute values using the

cross-sectional mean of available prices on that date. Fifth, we remove duplicate ISIN-date observations to ensure panel uniqueness. Following these filters, the cleaned dataset comprises 265,792 daily observations across 108 unique securities. Extreme returns are not deleted but are treated through winsorization, whose lower and upper bounds are tuned within the cross-validation loop (Section 3.8); this limits the influence of data-entry errors and thin-trading price jumps while preserving the genuine heavy tails characteristic of frontier-market returns.

3.3 Feature Engineering

We construct an extensive predictor set from daily OHLCV data, subsequently aggregated to weekly frequency to match our forecasting horizon. The final feature matrix contains 1,108 predictors organized into six thematic categories, each motivated by distinct theoretical or empirical rationales.

The first category comprises explosive behavior diagnostics based on recursive right-tailed unit root tests designed to detect and date-stamp periods of explosive price dynamics (Phillips, Shi, and Yu 2015; Vasilopoulos, Pavlidis, and Martínez-García 2022). The second category extracts outputs from univariate forecast models. Specifically, we obtain point forecasts and prediction intervals from three automated univariate specifications: ARIMA, exponential smoothing (ETS), and neural network autoregression (NNETAR), implemented via the forecast package in R (Hyndman and Khandakar 2008). Rather than using these forecasts directly as trading signals, we treat them as derived features capturing different aspects of time-series momentum and mean reversion.

The third category captures structural break indicators through backward CUSUM statistics that detect and quantify structural instability in return dynamics (Otto and Breitung 2023), computed using the backCUSUM package in R. The fourth category encompasses canonical time-series characteristics. We extract 22 canonical time-series characteristics selected for computational efficiency and discriminative power using the catch22 package in R (Lubba et al. 2019). These features capture autocorrelation structure, distributional properties, entropy measures, and nonlinear dynamics. We supplement these with additional features from the feasts and tsfeatures packages in R (O'Hara-Wild et al. 2023) and the TSFEL library in Python (Barandas et al. 2020).

The fifth category consists of wavelet-based features derived by applying discrete wavelet transformations to decompose price series into multi-resolution components, then generating forecasts from wavelet-ARIMA hybrid models using the WaveletArima package in R (Aminghafari and Poggi 2007; Paul, Samanta, and Yeasin 2022). The sixth and final category comprises technical indicators and market microstructure variables, computed using the finfeatures package in R (Sagovac 2023). We compute a comprehensive suite

including momentum oscillators (RSI, MACD, stochastics), trend indicators (moving average crossovers, ADX), volatility measures (Bollinger Bands, ATR), and volume-based indicators (OBV, VWAP).

Following feature generation, predictors with greater than 50 percent missing values, greater than 2 percent infinite values, or constant values are excluded. The final analytical sample comprises 67,806 weekly stock-level observations and 1,108 predictors.

From this filtered candidate set we retain, within each training fold, the 25 top-ranked predictors selected by an information-theoretic filter. Whether that filter is Joint Mutual Information (JMI) or ReliefF (Kononenko 1994) is itself tuned on the validation set: both rank predictors by their incremental information about the target while penalizing redundancy among already-selected features. The feature ranking is re-estimated inside each training window using only past data, so no future information enters the selected feature set. Restricting each model to 25 predictors keeps the learners well-conditioned given the low signal-to-noise ratio of weekly frontier-market returns.

3.4 Target Variable

The prediction target is the weekly simple return, computed as

$$r_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}$$

where $P_{i,t}$ denotes the closing price of security i at the end of week t and $r_{i,t}$ is its weekly simple return. We use simple rather than log returns to facilitate direct economic interpretation of portfolio performance.

The distribution of weekly returns across the pooled sample exhibits near-zero mean and median (0.00), moderate dispersion ($\sigma = 7\%$), and pronounced non-normality. Positive skewness (0.24) indicates a slight rightward tail asymmetry, while extreme kurtosis (18.30) signals heavy tails substantially exceeding the Gaussian benchmark. The return range spans from -42% to $+46\%$, reflecting the heightened volatility characteristic of frontier equity markets.

3.5 Model Specification

We evaluate four classes of supervised learning algorithms spanning the complexity spectrum from penalized linear models to nonlinear ensemble methods. This diversity enables assessment of whether predictive gains derive from capturing linear factor structure or exploiting nonlinear interactions (Gu, Kelly, and Xiu 2020). All models are implemented

as regression tasks predicting continuous weekly returns, though we also evaluate classification performance by dichotomizing predictions at zero. The four classes are chosen to span the complexity ladder used in the benchmark asset-pricing machine-learning literature: the penalized linear model isolates linear factor structure, the tree ensembles capture threshold effects and predictor interactions with minimal tuning, and the neural network probes for smooth nonlinearities, so that performance differences across the ladder are themselves informative about the functional form of any predictability.

Elastic Net (GLMNET) serves as our baseline, providing a benchmark against which nonlinear methods must demonstrate incremental value. We implement the Elastic Net estimator (Zou and Hastie 2005), which nests both LASSO and Ridge regression as special cases. The regularization parameter λ controls overall regularization strength while α governs the balance between L1 (sparsity-inducing) and L2 (shrinkage) penalties. The L1 component performs implicit variable selection by driving coefficients exactly to zero, while the L2 component stabilizes estimation when predictors exhibit multicollinearity.

Random Forest (Breiman 2001) constructs an ensemble of B regression trees, each trained on a bootstrap sample of the data with a random subset of m predictors considered at each split. By decorrelating trees through feature subsampling, Random Forest reduces the average pairwise correlation between tree predictions and achieves variance reduction beyond what bagging alone provides (Biau and Scornet 2016). This property is particularly valuable in our setting, where the low signal-to-noise ratio makes variance control paramount.

Gradient Boosted Trees (XGBoost) builds trees sequentially, with each tree fitted to the pseudo-residuals of the cumulative model (Friedman 2001). We implement XGBoost (Chen and Guestrin 2016), which augments standard gradient boosting with regularization terms that penalize tree complexity. This explicit complexity control, combined with column subsampling and shrinkage, makes XGBoost particularly resistant to overfitting in high-dimensional settings.

Feedforward Neural Network is a shallow architecture with a single hidden layer to assess whether flexible nonlinear function approximation improves upon tree-based methods. We deliberately limit network depth, as deeper architectures require substantially more data to avoid overfitting and have shown limited incremental benefit for tabular financial data relative to gradient boosted trees (Grinsztajn, Oyallon, and Varoquaux 2022). Weight decay regularization provides additional overfitting control.

Forecast Combination aggregates predictions from all four base models. We construct mean and median ensemble forecasts. The mean combination exploits potential complementarities across model classes, while the median provides robustness to outlier predictions (Timmermann 2006).

3.6 Nested Rolling-Window Cross-Validation

Validating predictive models on financial time series requires careful attention to temporal ordering. Standard k-fold cross-validation, which randomly partitions observations, violates the sequential structure of returns and permits information leakage from future to past (Gu, Kelly, and Xiu 2020). We therefore implement nested rolling-window cross-validation (NRWCV), a two-layer procedure that simultaneously provides unbiased performance estimates and optimizes hyperparameters without contamination.

The outer loop handles performance estimation. A rolling window advances through the sample chronologically, at each step training the model on historical data and evaluating predictions on a held-out future period. This design mimics the real-time forecasting environment that practitioners face. Nested within this outer loop, an inner loop conducts hyperparameter tuning by further partitioning each outer-loop training set into training and validation subsets using the same rolling-window logic. This nested structure ensures that the test set remains entirely untouched during hyperparameter search. To guard against information leakage from overlapping return calculations or autocorrelated features, we impose one-week gaps between training and validation sets, as well as between validation and test sets.

Table 1. *Nested rolling-window (expanding-origin) cross-validation window lengths, in weeks. Each row corresponds to one of the two cross-validation configurations; columns give the training, embargo gap, validation, second embargo gap, and test window lengths applied at each step of the scheme.*

Training	Gap	Validation	Gap	Test
192	1	12	1	1
288	1	24	1	1

Source: Authors' calculations.

We examine two parameterizations. The short-horizon configuration allocates 192 weeks (approximately four years) to training, 12 weeks to validation, and one week to testing. The long-horizon configuration extends these windows to 288 weeks (approximately six years) for training and 24 weeks for validation. These configurations embody a fundamental tradeoff: the short-horizon specification prioritizes adaptability to regime changes, while the long-horizon specification provides more stable parameter estimates. To be explicit, the validation protocol is a rolling-window (expanding-origin) scheme rather than k-fold cross-validation, which would be invalid for time-series data because random folds permit look-ahead leakage. The window lengths for each stage are reported in Table 2, and one-week embargo gaps are imposed between the training, validation, and test blocks to prevent leakage from overlapping return windows and autocorrelated features.

3.7 Performance Evaluation

We evaluate predictive performance through complementary statistical and economic lenses, addressing the well-documented gap between forecast accuracy and investment profitability (Leitch and Tanner 1991).

Statistical Metrics. For continuous return predictions, we report mean squared error (MSE) and mean absolute error (MAE). Although our models produce continuous forecasts, we dichotomize predictions at zero and evaluate binary classification performance through accuracy, precision, recall (TPR), and F-beta score.

Economic Evaluation. We conduct a portfolio backtest simulating a realistic investment strategy. Each week t, we form an equal-weighted long-only portfolio comprising all securities with positive predicted returns. We evaluate cumulative return, maximum drawdown, and Sharpe ratio (annualized excess return divided by annualized volatility). We report the baseline backtest gross of transaction costs; because realistic frictions on the ZSE are large and highly heterogeneous across the liquidity spectrum, we treat their impact as a separate sensitivity question (discussed in Sections 4.4 and 4.6) rather than embedding a single ad hoc cost assumption in the headline results. Portfolios are equal-weighted rather than mean-variance optimized: in a market this thin the sample covariance matrix is severely ill-conditioned, and equal weighting avoids compounding forecast error with the covariance-estimation error that destabilizes mean-variance optimization when few, thinly traded assets are available.

3.8 Hyperparameter Optimization

Hyperparameters are optimized via random search over predefined search spaces within the inner cross-validation loop, ensuring that hyperparameter selection does not contaminate out-of-sample evaluation. Within the inner loop, each candidate hyperparameter configuration is scored by validation-set mean squared error, and the configuration that minimizes it is selected and then refit on the combined training-plus-validation window before generating test-set forecasts. The same out-of-sample protocol governs the preprocessing hyperparameters described below, so that the winsorization bounds, the correlation-filter threshold, and the feature-selection method are all chosen without reference to the test set.

Preprocessing Hyperparameters. Winsorization bounds control treatment of extreme returns (lower bounds 0.001 to 0.2; upper bounds 0.8 to 0.999). The correlation filter threshold determines aggressiveness of multicollinearity reduction (0.80, 0.90, 0.95, 0.99). Feature selection method choice compares JMI and ReliefF.

Model Hyperparameters. For Random Forest, we optimize maximum tree depth (1–15),

feature subsample ratio (0.1–1.0), number of trees (10–2000), and bootstrap sampling method. For XGBoost, the search space spans learning rate (0.0001–1.0), maximum depth (1–20), boosting rounds (1–5000), L1 regularization (0.001–100), and row subsampling (0.1–1.0). For the neural network, we optimize hidden layer size (2–15 units), weight decay (0.0001–0.1), and maximum iterations (50–500). For Elastic Net, we optimize the mixing parameter alpha (0.0001–1.0) and regularization path index (5–30).

All preprocessing transformations are computed exclusively on training data within each cross-validation fold and applied out-of-sample to validation and test sets.

4 Results

This section presents the empirical findings in two parts. The first part (Section 4.1) examines statistical forecast performance across models and cross-validation configurations. The second part (Sections 4.2 through 4.9) evaluates economic significance through portfolio backtests, progressively examining aggregate portfolio performance, the liquidity gradient in predictability, transaction-cost implications, model heterogeneity across liquidity strata, the statistical significance of forecast and portfolio differences, robustness to transaction costs and rebalancing frequency, model interpretability, and an overall discussion of mechanisms.

4.1 Statistical Forecast Performance

Table 3 reports classification and regression metrics for all models under both cross-validation configurations. Several patterns emerge from these results.

Table 2. *Out-of-sample statistical forecast performance by model and cross-validation configuration. Classification metrics — directional accuracy, F-beta, true-positive rate (TPR), precision, true-negative rate (TNR), and negative predictive value (NPV) — use the sign of the predicted return as the class label; regression metrics are mean squared error (MSE) and mean absolute error (MAE) on continuous return predictions. The CV column reports the number of out-of-sample test windows.*

CV	Model	Accuracy	F.beta	TPR	Precision	TNR	NPV	MSE	MAE
1057	glmnet	0.48	0.55	0.72	0.45	0.28	0.56	0.01	0.04
949	glmnet	0.46	0.56	0.78	0.44	0.21	0.56	0.01	0.04
1057	nnet	0.53	0.53	0.60	0.48	0.48	0.60	0.04	0.04
949	nnet	0.53	0.53	0.60	0.47	0.47	0.60	0.02	0.04
1057	ranger	0.52	0.55	0.66	0.47	0.41	0.60	0.01	0.04
949	ranger	0.52	0.55	0.66	0.47	0.42	0.61	0.01	0.04
1057	xgboost	0.50	0.56	0.71	0.46	0.33	0.59	0.11	0.06

949	xgboost	0.50	0.55	0.70	0.46	0.35	0.60	10.65	0.14
1057	mean_resp	0.51	0.57	0.72	0.47	0.34	0.61	0.01	0.05
949	mean_resp	0.51	0.56	0.72	0.46	0.35	0.61	0.67	0.07
1057	median_resp	0.52	0.57	0.71	0.47	0.36	0.61	0.01	0.04
949	median_resp	0.51	0.56	0.71	0.46	0.36	0.61	0.00	0.04

Source: Authors' calculations.

The most striking observation is that all models achieve only modest directional accuracy, ranging from 46 percent (Elastic Net, CV-949) to 53 percent (Neural Network). These figures barely exceed the 50 percent threshold expected from random guessing, confirming the well-documented difficulty of return prediction and the low signal-to-noise environment characteristic of equity markets (Campbell and Thompson 2008). The neural network achieves the highest accuracy, though the margin over tree-based methods is economically modest. Notably, the penalized linear model consistently underperforms nonlinear alternatives, suggesting that return predictability on the ZSE derives at least partially from nonlinear predictor interactions that linear specifications cannot capture. This finding aligns with the broader machine learning literature documenting the importance of capturing interaction effects and nonlinear relationships in financial data.

The decomposition of classification performance into precision and recall reveals an important asymmetry. All models exhibit high true positive rates (60 to 78 percent) but substantially lower precision (44 to 48 percent). This pattern indicates a systematic bias toward predicting positive returns, which has implications for portfolio construction. A model with high recall but low precision will correctly identify most positive return opportunities but will also generate many false signals, potentially diluting portfolio performance. Table 4 confirms that models classify 56 to 75 percent of observations as positive. Elastic Net displays the most extreme imbalance, predicting positive returns for 75 percent of observations despite the near-zero unconditional mean return. This positive bias likely reflects the slight right-skewness of the return distribution and the asymmetric costs implicit in the squared error loss function used during training.

The F-beta scores, which balance precision and recall, show less differentiation across models (range: 0.53–0.57), with ensemble methods and XGBoost marginally outperforming the neural network. This divergence from the accuracy ranking underscores that model selection depends critically on the practitioner's loss function. Investors penalizing false positives more heavily than missed opportunities would prefer models with higher precision despite lower overall accuracy.

Table 3. *Share of positive versus negative weekly-return predictions by model (percent of all out-of-sample predictions). A high positive share indicates a model that predominantly forecasts price increases.*

Model	Negative Share	Positive Share
glmnet	24.86	75.14
nnet	44.20	55.80
ranger	37.99	62.01
xgboost	32.25	67.75
mean_resp	31.69	68.31
median_resp	32.62	67.38

Source: Authors' calculations.

Comparing results across cross-validation configurations reveals reassuring stability. Performance metrics differ only marginally between CV-949 and CV-1057, suggesting that our findings are not artifacts of a particular temporal parameterization.

4.2 Portfolio Backtest Results

Statistical forecast accuracy does not directly address economic significance. We evaluate models through a portfolio backtest that simulates a realistic investment strategy.

Table 4. *Annualized performance of the equal-weighted, long-only portfolio by model and cross-validation configuration. Columns report annualized return, annualized standard deviation (SD), the annualized Sharpe ratio, maximum peak-to-trough drawdown, and the Sortino ratio. Results use the full security universe and are gross of transaction costs.*

CV	Model	Annual Return	Annual SD	Sharpe	Max Drawdown	Sortino
1057	glmnet	0.12	0.17	0.70	0.78	0.15
1057	nnet	0.44	0.19	2.26	0.75	0.45
1057	ranger	0.39	0.21	1.90	0.73	0.39
1057	xgboost	0.31	0.19	1.59	0.72	0.32
1057	mean_resp	0.41	0.19	2.18	0.69	0.43
1057	median_resp	0.45	0.20	2.29	0.74	0.46
949	glmnet	0.04	0.17	0.25	0.79	0.06
949	nnet	0.30	0.18	1.68	0.64	0.33
949	ranger	0.30	0.19	1.58	0.69	0.31
949	xgboost	0.23	0.19	1.25	0.74	0.25
949	mean_resp	0.30	0.18	1.65	0.72	0.32
949	median_resp	0.31	0.19	1.66	0.73	0.32

Source: Authors' calculations.

Table 5 reports annualized portfolio statistics. The equal-weighted long-only strategy generates substantial returns across most specifications, with annualized performance ranging from 4 percent (Elastic Net, CV-949) to 45 percent (median ensemble, CV-1057). These results suggest that the modest statistical predictability documented in the previous section does translate into economically meaningful portfolio returns, though the magnitude varies considerably across model specifications.

Elastic Net consistently underperforms, achieving Sharpe ratios of 0.25 and 0.70, below typical hurdle rates for active strategies. This underperformance is consistent with the model's high false positive rate. By predicting positive returns for 75 percent of observations, Elastic Net dilutes portfolio quality by including many securities with near-zero or negative expected returns. The lesson for practitioners is clear: statistical significance does not guarantee economic value, and models with systematic directional biases may underperform despite reasonable accuracy metrics.

Tree-based methods deliver substantially stronger risk-adjusted performance. Random Forest achieves Sharpe ratios of 1.58 and 1.90, while XGBoost produces 1.25 and 1.59. These figures represent economically meaningful outperformance, corresponding to annualized returns of 30 to 39 percent against volatility of approximately 19 to 21 percent. The neural network performs comparably, with Sharpe ratios of 1.68 and 2.26, the latter representing the highest single-model performance in the CV-1057 configuration. Ensemble methods match or exceed the best individual models, with the median ensemble achieving 1.66 and 2.29. That simple averaging of diverse model predictions improves upon constituents is consistent with the forecast combination literature and suggests that individual models capture complementary aspects of return predictability.

Table 5. *Average annualized Sharpe ratio by model, averaged across the two cross-validation configurations (full universe, gross of transaction costs).*

Model	Sharpe Ratio
glmnet	0.48
xgboost	1.42
ranger	1.74
mean_resp	1.91
nnet	1.97
median_resp	1.97

Source: Authors' calculations.

Table 6 summarizes average Sharpe ratios, confirming that nonlinear methods substan-

tially outperform the linear benchmark and that forecast combination provides incremental value.

Maximum drawdown statistics merit attention. All strategies experience severe peak-to-trough declines ranging from 64 to 79 percent, levels that would test the risk tolerance of most investors. These drawdowns likely concentrate during the 2008 financial crisis and March 2020 COVID-19 shock, periods when long-only equity strategies suffered globally. The similarity of drawdowns across models suggests none possesses meaningful defensive characteristics during market stress. This finding has important implications for portfolio construction: even if machine learning models generate attractive average returns, risk management tools beyond model selection are needed to navigate crisis periods.

4.3 The Liquidity Gradient in Predictability

Our most novel finding concerns the relationship between stock liquidity and forecast performance. Tables 7 and 8 report Sharpe ratios when the investment universe is restricted to the n most liquid securities. We rank all sample securities by their average weekly trading turnover and define a sequence of nested liquidity universes (or strata): the 10, 20, 30, and 40 most liquid securities, together with the full universe ("All"). A smaller universe therefore corresponds to a more liquid, more actively traded subset, while widening the universe progressively admits thinner, less frequently traded stocks.

Table 6. *Annualized Sharpe ratio by model across nested liquidity universes. Columns restrict the investable universe to the 10, 20, 30, and 40 most-traded securities (ranked by average weekly turnover) and to the full universe (All); gross of transaction costs.*

Model	Top 10	Top 20	Top 30	Top 40	All
glmnet	0.18	0.24	0.25	0.30	0.48
nnet	0.17	0.37	0.58	0.87	1.97
ranger	0.09	0.24	0.55	0.82	1.74
xgboost	0.34	0.49	0.60	0.75	1.42
mean_resp	0.16	0.34	0.58	0.86	1.91
median_resp	0.06	0.25	0.44	0.83	1.97

Source: Authors' calculations.

Table 7. *Average annualized Sharpe ratio by liquidity-universe size n (the number of most-liquid securities included), averaged across all models.*

n	Sharpe Ratio
10	0.17

20	0.32
30	0.50
40	0.74
All	1.58

Source: Authors' calculations.

The results reveal a striking monotonic pattern. Predictability increases systematically as the universe expands to include less liquid securities. Averaging across models, the Sharpe ratio rises from 0.17 when trading only the 10 most liquid stocks to 1.58 when including all securities, a nearly tenfold increase. This gradient is economically large and statistically consistent across model specifications.

The pattern holds within each model class, though magnitudes vary. XGBoost exhibits the flattest gradient, achieving a Sharpe ratio of 0.34 on the top-10 liquid universe versus 1.42 on the full universe (ratio: 4.2x). The neural network and median ensemble show the steepest gradients, rising from approximately 0.10–0.17 on liquid stocks to 1.97 on the full universe (ratio: 12–20x). This heterogeneity suggests that tree-based methods may better capture predictable patterns in liquid securities, while neural networks and ensembles excel at exploiting inefficiencies in the illiquid tail.

Three mechanisms, not mutually exclusive, may explain this liquidity gradient. First, illiquid securities may command a liquidity risk premium that manifests as higher average returns (Amihud 2002). If models partially capture liquidity-related return variation, the premium would inflate backtest performance without reflecting genuine forecasting skill. Second, less liquid securities receive lower analyst coverage and institutional attention, potentially permitting mispricings that persist long enough for weekly rebalancing strategies to exploit. Third, and most critically, our backtest abstracts from transaction costs that would disproportionately erode returns on illiquid positions.

4.4 Transaction Cost Implications

The omission of transaction costs represents a significant limitation that likely overstates achievable performance, particularly for strategies emphasizing illiquid securities. Trading frictions on the ZSE include explicit costs (brokerage commissions of approximately 0.3 to 0.5 percent per transaction) and implicit costs (bid-ask spreads, market impact) that are substantially higher for illiquid names.

Bid-ask spreads on illiquid ZSE securities can exceed 2 to 5 percent, implying that the superior backtest performance on the full universe would be substantially and perhaps entirely eroded by implementation costs. The finding that predictability concentrates in

illiquid securities may therefore reflect an illiquidity illusion rather than exploitable inefficiency (Lesmond, Ogden, and Trzcinka 1999).

4.5 Model Heterogeneity Across Liquidity Strata

Figure 2 displays cumulative equity curves for all models in the full security universe. The liquidity-strata results are reported numerically in Tables 7 and 8; the figure is used to visualize time variation in aggregate strategy performance.

> **Note**
>
> *Figure placeholder — add* `fig4_cropped.png` *to the* `paper/` *folder and re-render to display the chart.*

Figure 2. *Cumulative gross portfolio value of the equal-weighted long-only strategy across nested liquidity strata (CV-1057). Horizontal axis: calendar time over the out-of-sample period. Vertical axis: cumulative portfolio value indexed to 1 at the start of the backtest.*

For the most liquid universe (top-10 securities), XGBoost dominates alternatives, consistent with its relatively flat liquidity gradient. The equity curves for other models are largely flat or declining, suggesting that machine learning methods struggle to identify exploitable patterns among the most efficient ZSE securities. This finding aligns with efficient market intuitions: the most liquid, widely followed stocks should exhibit the least predictability.

As the universe expands, model rankings shift. At 30–40 securities, Random Forest and ensemble methods emerge as top performers, while the neural network gains prominence only when the full universe is included. This pattern suggests that different model architectures capture distinct types of predictability. Tree-based methods may excel at exploiting well-defined technical patterns in moderately liquid securities, while neural networks identify subtler nonlinear relationships in the illiquid tail.

The equity curves also reveal substantial time-variation in strategy performance. All models suffer significant drawdowns during the 2008 financial crisis and experience performance degradation in certain subperiods. This instability underscores the importance of realistic expectations: even successful forecasting models will experience extended periods of underperformance that may challenge investor commitment.

4.6 Statistical Significance of Forecast and Portfolio Differences

The cross-model comparisons in Sections 4.1–4.3 are tested formally below. Forecast accuracy is assessed with a Diebold-Mariano test applied to the weekly squared-error loss

differential between each model and the Elastic Net benchmark. Economic performance is assessed with stationary block-bootstrap 95 percent confidence intervals for the annualized Sharpe ratio and with pairwise bootstrap tests of the Sharpe-ratio difference relative to the benchmark.

Table 8. *Diebold-Mariano tests of out-of-sample forecast accuracy, each model versus the Elastic Net benchmark, by cross-validation configuration. The loss differential is the difference in weekly squared forecast errors. A negative statistic indicates that the model has lower mean squared error than the benchmark; p-values are two-sided.*

Model	DM (CV-949)	p (CV-949)	DM (CV-1057)	p (CV-1057)
Mean Ensemble	1.005	0.3151	1.833	0.0670
Median Ensemble	-1.886	0.0596	-2.630	0.0087
Neural Network	1.061	0.2888	1.264	0.2063
Random Forest	1.767	0.0776	2.701	0.0070
XGBoost	1.003	0.3161	1.403	0.1610

Source: Authors' calculations.

Table 9 shows that forecast-error differences relative to Elastic Net are model- and window-dependent. The clearest lower-loss result is observed for Median Ensemble in CV-1057, where the negative statistic indicates lower weekly mean squared forecast loss than the benchmark. A weaker, marginal lower-loss result is observed for Median Ensemble in CV-949. By contrast, Random Forest in CV-1057 is statistically significant in the opposite direction, indicating higher squared-error loss than Elastic Net despite stronger portfolio performance. This qualifies H2: nonlinear and ensemble models are economically stronger in the backtest, but their one-week squared-error advantage is not uniform across model classes.

This divergence between the statistical and the economic criteria is itself a substantive result rather than a complication. The Diebold-Mariano tests speak to squared-error loss, which in weekly return data is dominated by the unpredictable magnitude of price changes; the portfolio backtest, by contrast, rewards the directional and cross-sectional ranking signal that a long-only strategy actually exploits. That almost no model lowers squared-error loss relative to Elastic Net while every nonlinear model raises the portfolio Sharpe ratio is precisely the wedge between forecast accuracy and economic value documented by Leitch and Tanner (1991), and it vindicates the dual-evaluation design adopted in Section 3.7. Notably, the Median Ensemble is the only specification that improves on the benchmark on both criteria at once — a lower squared-error loss (Table 9) and the highest risk-adjusted return (Table 10) — which identifies forecast combination as the most robust specification and motivates its use in the robustness analysis of Section 4.7.

Table 9. *Annualized Sharpe ratios with stationar*
cross-validation configuration (full uni
difference in Sharpe ratio relative to the El
2,000 b

CV / Model	Sharpe	95% CI lower	95% CI upper	Delta vs Elastic Net
CV-949 / Elastic Net	*0.338*	*-0.301*	*1.070*	—
CV-949 / Random Forest	1.486	0.765	2.338	1.148
CV-949 / XGBoost	1.221	0.462	2.077	0.883
CV-949 / Neural Network	1.559	0.848	2.402	1.221
CV-949 / Mean Ensemble	1.536	0.796	2.418	1.198
CV-949 / Median Ensemble	1.544	0.759	2.433	1.206
CV-1057 / Elastic Net	*0.747*	*0.078*	*1.595*	—
CV-1057 / Random Forest	1.714	1.017	2.453	0.966
CV-1057 / XGBoost	1.488	0.779	2.279	0.741
CV-1057 / Neural Network	1.981	1.203	2.825	1.233
CV-1057 / Mean Ensemble	1.930	1.188	2.765	1.182
CV-1057 / Median Ensemble	1.995	1.267	2.772	1.248

Source: Authors' calculations.

[1] The point estimates reported here reflect a strictly aligned common-window sample necessary for valid

Table 10 reports annualized Sharpe ratios with stationary-block-bootstrap 95 percent confidence intervals, separately by cross-validation configuration. All 10 nonlinear and ensemble specifications have confidence intervals above zero; Elastic Net does so only in the CV-1057 configuration. Sharpe-difference tests indicate statistically detectable outperformance over Elastic Net for all 10 nonlinear and ensemble specifications.

Two cautions temper these comparisons. First, the bootstrap confidence intervals for the individual Sharpe ratios are wide and substantially overlapping across the nonlinear and ensemble models — the neural network, random forest, and median ensemble intervals all intersect — so the data support the claim that the class of nonlinear and ensemble models outperforms the linear benchmark but do not support ranking these models reliably against one another. Second, the Elastic Net benchmark is itself statistically indistinguishable from zero skill in the CV-949 configuration, where its Sharpe-ratio confidence interval spans zero, underscoring how little of the predictability a purely linear specification captures.

4.7 Robustness Checks: Transaction Costs, Rebalancing Frequency, and Benchmark Comparison

To assess whether the headline results survive realistic implementation frictions and design choices, the backtest is re-run for the median-ensemble strategy under proportional transaction costs, lower monthly rebalancing frequency, and an explicit comparison with the official ZSE CROBEX price-index benchmark. Transaction costs are applied to portfolio turnover, including the initial purchase.

Table 10. *Robustness of median-ensemble portfolio performance to transaction costs, rebalancing frequency, and benchmark comparison (CV-1057, common official CROBEX overlap period). Columns report the annualized return, Sharpe ratio, Sortino ratio, and maximum drawdown under each specification. The CROBEX row is a passive buy-and-hold price-index benchmark.*

Specification	Ann. return	Sharpe	Sortino	Max drawdown
Baseline (gross, weekly rebalancing)	0.442	2.494	3.401	0.287
Net of costs (50 bps per trade)	0.157	1.057	1.459	0.391
Net of costs (150 bps per trade)	-0.257	-1.919	-2.561	0.986
Monthly rebalancing (gross)	0.114	0.817	1.003	0.343
Monthly rebalancing (net, 50 bps)	0.056	0.451	0.549	0.409
CROBEX buy-and-hold (benchmark)	**0.021**	**0.236**	**0.278**	**0.397**

Source: Authors' calculations.

Table 11 evaluates the median-ensemble strategy under implementation frictions and lower turnover, using the official ZSE CROBEX price index as the passive benchmark. CROBEX official ZSE overlap: 2010-01-16 to 2024-03-16 (740 weekly returns). The gross weekly Sharpe ratio is 2.49; applying 50 bps and 150 bps per trade lowers it to 1.06 and -1.92, respectively. Monthly rebalancing produces a Sharpe ratio of 0.82, showing how much performance remains when turnover is mechanically reduced. The CROBEX buy-and-hold benchmark has a Sharpe ratio of 0.24 over the same official-data overlap period.

These figures sharpen, rather than overturn, the economic interpretation. The gross weekly Sharpe ratio of 2.49 falls below zero between the 50- and 150-basis-point cost levels, implying a break-even round-trip cost on the order of one percent. This threshold interacts directly with the liquidity gradient of Section 4.3: the full-universe performance that produces the highest Sharpe ratios is generated by the least-liquid securities, for which bid-ask spreads of two to five percent (Section 4.4) lie well above the break-even cost, whereas the liquid securities for which a 50-basis-point assumption is realistic deliver only modest gross Sharpe ratios (Table 7). Risk-adjusted predictability is therefore largest precisely where it is most expensive to harvest, and cheapest to harvest precisely

where it is smallest — an implementability tension that the net-of-cost results quantify and that supports reading the headline magnitudes as an upper bound on achievable performance.

Two further patterns are informative. Lowering the rebalancing frequency from weekly to monthly reduces the gross Sharpe ratio from 2.49 to 0.82, indicating that most of the exploitable signal is short-lived and decays within the month; because capturing it requires near-complete weekly turnover, the strategy is structurally exposed to transaction costs and cannot escape them simply by trading less. A genuine but more modest edge nonetheless survives under favourable yet defensible assumptions: net of 50 basis points the median ensemble retains a Sharpe ratio of 1.06, roughly four times the contemporaneous CROBEX buy-and-hold value of 0.24, and even monthly net-of-cost rebalancing (0.45) exceeds the passive benchmark. The Sortino ratio exceeds the Sharpe ratio in every specification, indicating that the strategy's variability is weighted toward upside rather than downside deviations. The contribution of this study is thus best read as evidence on the anatomy of return predictability in a frontier market — its existence, its nonlinear structure, and its concentration in illiquid securities — rather than as a directly tradable strategy.

4.8 Model Interpretability

To provide economic intuition for which signals drive the forecasts, a feature-importance analysis is reported for a representative specification. The production results were generated from a large nested rolling-window design in which fitted learner objects were not serialized to keep the experiment within its compute and storage budget. Feature attribution is therefore not recovered from the pooled production models themselves. The archived results nonetheless preserve the out-of-sample predictions, the selected hyperparameters, the retained feature names, and the random seeds. For the interpretability analysis, a single gradient-boosted-tree model is re-estimated on a representative full-universe training window using the identical preprocessing and Joint Mutual Information / ReliefF feature-selection pipeline, and permutation importance is aggregated to the level of the six predictor families defined in Section 3.3.

Table 12 reports each family's share of total permutation importance.

Table 11. *Share of total permutation importance by predictor family, for a gradient-boosted-tree model fitted on a representative full-universe training window. The importance of each predictor is the mean increase in out-of-sample mean squared error when that predictor is randomly permuted; importances are summed within each of the six predictor families of Section 3.3 and expressed as a percentage of the total. Because the number of predictors surviving selection differs across families, the table reports each family's share of total importance rather than a raw count. The analysis is illustrative of a representative training window and is not an attribution of the pooled production models.*

Predictor family	Share of total permutation importance (%)
Technical indicators and momentum	34.27
Volatility measures	23.81
Univariate forecast outputs (ARIMA / ETS / NNETAR)	16.54
Canonical time-series characteristics (catch22)	11.92
Wavelet-based features	8.15
Explosive-behaviour and structural-break diagnostics	5.31

Source: Authors' calculations.

Two features of this profile are anticipated. First, importance is expected to be diffuse rather than concentrated in any single predictor, consistent with the low signal-to-noise environment documented throughout Section 4 and with the interpretation that predictability arises from the combination of many weak signals rather than from a single dominant factor. Second, the families expected to carry the most weight are the economically interpretable ones — short-horizon momentum and volatility — aligning with the mean-reversion and volatility-timing mechanisms discussed in Section 4.9. The leading families are Technical indicators and momentum and Volatility measures, which together account for 58.08 percent of total importance. For practitioners this offers a parsimonious reading: most of the exploitable structure is captured by standard technical and volatility signals, while the more exotic high-dimensional features provide incremental refinement rather than the bulk of the predictive content.

4.9 Discussion: Mechanisms, Overfitting, and Data-Snooping

Two features of our design plausibly explain why the nonlinear framework outperforms the linear benchmark and a passive market position. First, the predictor set is deliberately high-dimensional and heterogeneous, combining momentum, volatility, structural-break, and time-series-feature signals; tree ensembles and the neural network can exploit interactions and threshold effects among these signals that the Elastic Net, restricted to linear combinations, cannot represent. Second, the gains concentrate in less-liquid securities, consistent with the view that slower information diffusion in thinly traded stocks leaves more exploitable structure for flexible learners to capture.

Several aspects of the design guard against overfitting. The nested rolling-window scheme selects all hyperparameters out-of-sample; regularization is intrinsic to every model class (the L1/L2 penalties of the Elastic Net, the depth limits, subsampling, and shrinkage of the tree ensembles, and weight decay for the neural network); and one-week embargo gaps prevent leakage from autocorrelated features. Nevertheless, the modest directional accuracy (46 to 53 percent) is a reminder that the signal-to-noise ratio is low and that apparent performance should be interpreted cautiously.

Because we evaluate several model classes, two cross-validation configurations, and five liquidity universes, the study is exposed to data-snooping risk: the best-looking specification may owe part of its performance to multiple testing (Sullivan, Timmermann, and White 1999). We mitigate this by reporting results for all specifications rather than a single hand-picked one, and by emphasizing the monotonic liquidity pattern that holds across every model rather than any single point estimate. Formal forecast-significance tests and bootstrap confidence intervals for the Sharpe ratios are reported in Section 4.6; the magnitudes are nonetheless interpreted cautiously, since the low weekly signal-to-noise ratio leaves some cross-model differences statistically indistinguishable.

Finally, the economic interpretation of our results hinges on transaction costs. As Section 4.4 details, the very securities that drive the headline performance are those for which bid-ask spreads and market-impact costs are largest, so net-of-cost returns are materially lower and the liquidity gradient may partly reflect an illiquidity illusion (Lesmond, Ogden, and Trzcinka 1999) rather than fully exploitable inefficiency. Section 4.7 reports a first net-of-cost and rebalancing-frequency robustness check against the CROBEX benchmark; extending this analysis across all liquidity strata and adding model-attribution evidence remains the principal agenda for future work.

5 Conclusion

This paper investigates whether machine learning methods can forecast weekly equity returns on the Zagreb Stock Exchange, a frontier market characterized by thin trading, concentrated sector composition, and limited institutional participation. Using daily OHLCV data spanning 2000–2024, we construct over 1,100 predictors encompassing technical indicators, time-series characteristics, explosive behavior diagnostics, and wavelet-based features. We evaluate four model classes within a nested rolling-window cross-validation framework designed to prevent information leakage and provide unbiased out-of-sample performance estimates.

Our findings contribute to the machine learning literature in asset pricing along several dimensions. First, we document modest out-of-sample return predictability on the ZSE.

Directional accuracy ranges from 46 to 53 percent across models, marginally exceeding random chance but consistent with the low signal-to-noise environment documented in developed market studies. Second, nonlinear methods substantially outperform the penalized linear benchmark in portfolio terms, suggesting that return predictability derives at least partially from complex predictor interactions that linear specifications cannot capture. Neural networks and forecast-combination ensembles achieve the highest overall Sharpe ratios, though formal forecast-error tests show that model differences are not uniformly precise at the weekly horizon.

The paper's most novel contribution concerns the relationship between liquidity and predictability. We document a pronounced monotonic gradient: risk-adjusted portfolio performance increases systematically as the investment universe expands to include less liquid securities, with Sharpe ratios rising from 0.17 for the ten most liquid stocks to 1.58 for the full universe. This pattern is consistent across model specifications and cross-validation configurations. We attribute this gradient to three potentially overlapping mechanisms: compensation for liquidity risk premia, reduced market efficiency in securities with lower analyst coverage, and the omission of transaction costs that would disproportionately penalize illiquid-focused strategies. The finding that XGBoost exhibits the flattest liquidity gradient while neural networks show the steepest suggests that different model architectures capture distinct types of predictability, with potential implications for model selection in practitioner applications.

Two quantitative findings from this revision sharpen these conclusions. First, the cross-model differences are economically robust but statistically imprecise at the weekly horizon, so the evidence supports the superiority of the class of nonlinear and ensemble models over the linear benchmark rather than the selection of any single best model. Second, the net-of-cost analysis implies a break-even trading cost near one percent: the median ensemble retains a Sharpe ratio of 1.06 net of 50 basis points and continues to exceed the CROBEX benchmark, yet the illiquid securities that drive the highest gross performance carry spreads above this threshold, so the headline magnitudes are best interpreted as an upper bound on achievable performance.

Several limitations warrant acknowledgment and motivate directions for future research. Most significantly, our backtest can only approximate transaction costs, market impact, and short-sale constraints. Given the documented illiquidity of many ZSE securities, where bid-ask spreads can exceed 2 to 5 percent for thinly traded names, implementation frictions materially reduce the profits suggested by gross backtests. The finding that predictability concentrates in illiquid securities amplifies this concern: the securities where models perform best are precisely those where execution costs are highest. In addition, the original experiment registry did not store fitted model objects, preventing

defensible SHAP or permutation-importance analysis without a targeted re-estimation.

Additional extensions merit investigation. Our predictor set, while extensive, excludes fundamental accounting data, macroeconomic variables, and cross-sectional characteristics (such as size and book-to-market ratios) that have proven predictive in developed market studies. Incorporating these dimensions could enhance forecast performance and permit decomposition of predictability into factor-related and residual components. Alternative neural network architectures, including recurrent networks capable of modeling longer temporal dependencies and attention mechanisms that weight predictor relevance dynamically, represent promising methodological extensions. Finally, extending the analysis to other Central and Eastern European markets (Warsaw, Budapest, Prague, Ljubljana) would test whether our findings generalize across frontier market settings or reflect ZSE-specific characteristics.

From a practical standpoint, our results offer cautious guidance for quantitative investors considering machine learning applications in frontier markets. The documented predictability, while statistically significant, is modest in magnitude and concentrated in securities where implementation is most challenging. Practitioners should view backtest results as upper bounds on achievable performance and conduct rigorous transaction cost sensitivity analysis before deployment. The finding that simple forecast combination improves upon individual models suggests that model averaging represents a low-cost robustness enhancement regardless of the specific algorithms employed. The liquidity-dependent heterogeneity in model performance implies that optimal algorithm selection may vary with the target investment universe, a consideration absent from most model comparison studies focused on developed markets.

In conclusion, this study demonstrates that machine learning methods generate economically meaningful return predictability on the Zagreb Stock Exchange, extending the evidence base beyond the deep, liquid markets where most prior research concentrates. The pronounced liquidity gradient we document represents a novel empirical finding with implications for both asset pricing theory and quantitative investment practice. Whether this predictability reflects genuine market inefficiency exploitable after transaction costs, compensation for liquidity risk, or artifacts of backtest methodology remains an open question that future research should address through more realistic implementation simulations. Our findings suggest that frontier equity markets, despite their data limitations and implementation challenges, offer fertile ground for machine learning applications, but that the gap between backtest performance and realized returns may be wider than in developed market settings.

References

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets, 5*(1), 31–56.

Amihud, Y., and Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics, 17*(2), 223–249.

Aminghafari, M., and Poggi, J.-M. (2007). Forecasting time series using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing, 5*(5), 709–724.

Azevedo, V., Kaiser, G. S., and Mueller, S. (2023). Stock market anomalies and machine learning across the globe. *Journal of Asset Management, 24*(5), 376–405.

Barandas, M., et al. (2020). TSFEL: Time series feature extraction library. *SoftwareX, 11*, 100456.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond risk premia with machine learning. *Review of Financial Studies, 34*(2), 1046–1089.

Biau, G., and Scornet, E. (2016). A random forest guided tour. *Test, 25*(2), 197–227.

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

Cakici, N., and Zaremba, A. (2024). What drives stock returns across countries? *International Review of Financial Analysis, 96*, 103548.

Campbell, J. Y., and Thompson, S. B. (2008). Predicting excess stock returns out of sample. *Review of Financial Studies, 21*(4), 1509–1531.

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

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, 785–794.

Choi, J., et al. (2025). International stock return predictability. *Journal of Financial Economics* (forthcoming).

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance, 66*(4), 1047–1108.

Consoli, S., Reforgiato Recupero, D., and Saisana, M. (2021). *Data science for economics and finance*. Springer.

De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley and Sons.

Fama, E. F. (1970). Efficient capital markets. *The Journal of Finance, 25*(2), 383–417.

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

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo. *The Journal of Finance, 75*(3), 1327–1370.

Fischer, T., and Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research, 270*(2), 654–669.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models. *Journal of Statistical Software, 33*(1), 1–22.

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

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? *NeurIPS 2022*.

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

Hanauer, M. X., and Kalsbach, T. (2023). Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review, 55*, 101022.

Hyndman, R. J., and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*(3), 1–22.

Jensen, T. I., Kelly, B. T., Malamud, S., and Pedersen, L. H. (2023). Machine learning and the implementable efficient frontier. *Swiss Finance Institute Research Paper No. 22-63*.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. *ECML 1994*, 171–182.

Kumbure, M. M., et al. (2022). Machine learning techniques for stock market forecasting: A literature review. *Expert Systems with Applications, 197*, 116659.

Leitch, G., and Tanner, J. E. (1991). Economic forecast evaluation: Profits versus conventional error measures. *American Economic Review, 81*(3), 580–590.

Lesmond, D. A., Ogden, J. P., and Trzcinka, C. A. (1999). A new estimate of transaction costs. *Review of Financial Studies, 12*(5), 1113–1141.

Liu, Y., et al. (2023). Transformer-based models for multi-asset forecasting. *Journal of Financial Data Science, 5*(2), 45–67.

Lubba, C. H., et al. (2019). catch22: CAnonical Time-series CHaracteristics. *Data Mining and Knowledge Discovery, 33*(6), 1821–1852.

Manojlović, T., and Štajduhar, I. (2016). Predicting stock market trends using random

forests. *MIPRO 2015*, 1189–1193.

O'Hara-Wild, M., et al. (2023). *feasts: Feature extraction and statistics for time series*. R package.

Otto, S., and Breitung, J. (2023). Backward CUSUM for testing and monitoring structural change. *Econometric Theory, 39*(4), 659–692.

Paul, R. K., Samanta, S., and Yeasin, M. (2022). *WaveletArima: Wavelet-ARIMA model for time series forecasting*. R package.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238.

Pesaran, M. H., and Timmermann, A. (2000). A recursive modelling approach to predicting UK stock returns. *Economic Journal, 110*(460), 159–191.

Phillips, P. C. B., Shi, S., and Yu, J. (2015). Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500. *International Economic Review, 56*(4), 1043–1078.

Sagovac, M. (2023). *finfeatures: Financial feature engineering functions*. R package.

Šego, B., and Škrinjarić, T. (2019). Kvantitativna istraživanja Zagrebačke burze. *Ekonomski pregled, 69*(6), 655–743.

Sullivan, R., Timmermann, A., and White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance, 54*(5), 1647–1691.

Thanh Nhon, H., Do-Thi, N., and Nguyen-Trang, T. (2025). Predicting stock returns using machine learning combined with data envelopment analysis and automatic feature engineering: A case study on the Vietnamese stock market. *PLOS ONE, 20*(9), e0332154.

Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, Vol. 1, 135–196.

Vasilopoulos, K., Pavlidis, E., and Martinez-Garcia, E. (2022). exuber: Recursive right-tailed unit root testing with R. *Journal of Statistical Software, 103*(10).

Zhang, G. P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, 30*(4), 451–462.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B, 67*(2), 301–320.